

Verena Schröter

Null Subjects in Englishes

Topics in English Linguistics

Editors

Elizabeth Closs Traugott

Susan M. Fitzmaurice

Bernd Kortmann

Volume 102

Verena Schröter

Null Subjects in Englishes

A Comparison of British English and Asian Englishes

DE GRUYTER
MOUTON

ISBN 978-3-11-063343-6

e-ISBN (PDF) 978-3-11-064926-0

e-ISBN (EPUB) 978-3-11-064535-4

Library of Congress Control Number: 2019947121

Bibliographic information published by the Deutsche Nationalbibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available on the Internet at <http://dnb.dnb.de>.

© 2019 Walter de Gruyter GmbH, Berlin/Boston

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

To Alex and Mira, who have been so patient and understanding,
and continue to be my inspiration and guiding star.

Acknowledgements

This book is a revised version of my dissertation submitted at the University of Freiburg in 2017. It would have been inconceivable without the help and encouragement of people too numerous to do justice to all. Still, first and foremost I want to thank my supervisor Bernd Kortmann for his continued faith and support, his helpful feedback, and for introducing me to the world of linguistics. Uncounted hours were spent at the Lehrstuhl Kortmann, and through all this time, Melitta Cocan has been a stronghold of sanity. Warmest thanks go to my long time colleagues and partners in times good and demanding, providing intense discussion, valuable feedback, or much needed distraction over a coffee or lunch. For polishing the final draft, Elizabeth Closs Traugott was an immense and ever reliable help.

I am more grateful than I can express for the unwavering support I received from my friends and family, who believed in me when I couldn't. Without their help in finding my way, and especially the most loving and reliable childcare, this dissertation could not have been written. Special thanks go to my sister Vanessa for hosting me during a writing binge, and for her diligent and just in time proof-reading. All remaining errors are mine and mine alone.

Contents

Acknowledgements — VII

List of Tables — XIII

List of Figures — XVII

List of Abbreviations — XIX

1 Introduction — 1

- 1.1 Null subjects — 1
- 1.2 Research context and method — 3
- 1.3 Database — 5
- 1.4 Scope and structure of the present study — 8

Part A: Null subjects in British English

2 Null subjects: Theoretical, methodological and descriptive foundations — 15

- 2.1 Pro-forms and pro-drop — 15
 - 2.1.1 Pro-forms — 16
 - 2.1.2 Pronouns and information structure — 18
 - 2.1.3 Pro-drop — 23
- 2.2 Quantitative approaches — 27
 - 2.2.1 The variationist paradigm and grammatical variation — 27
 - 2.2.2 Variationist studies on null subjects — 29
- 2.3 Null subjects in L1 English — 34
 - 2.3.1 English as a non-null subject language? — 34
 - 2.3.2 Quantitative research on null subjects in English — 37

3 Empirical baseline: Null subjects in Spoken British English — 44

- 3.1 The envelope of variation: Coding decisions and examples — 44
- 3.2 Analytic categories: Linguistic factors — 47
 - 3.2.1 Coreferential coordination — 47
 - 3.2.2 Clause type — 49
 - 3.2.3 Position — 50
 - 3.2.4 Person — 51

3.2.5	Specific reference —	52
3.2.6	Switch reference —	53
3.2.7	Persistence —	54
3.2.8	Verb phrase —	55
3.3	Overview ICE Great Britain: Descriptive statistics and methodology —	57
3.3.1	Distribution of null subjects in the data set —	57
3.3.2	Extra-linguistic factors —	59
3.3.3	Collocation —	61
3.3.4	Summary: Subject tokens and linguistic factors —	64
3.3.5	Statistical method —	66
3.4	Null subjects in ICE Great Britain: Logistic regression model —	67
3.4.1	Model selection —	67
3.4.2	Model evaluation —	69
3.4.3	Results by factor groups —	72
3.4.3.1	Coreferential coordination (“and”) —	72
3.4.3.2	Clause —	74
3.4.3.3	Position —	75
3.4.3.4	Person —	76
3.4.3.5	Specific reference —	77
3.4.3.6	Switch reference —	78
3.4.3.7	Persistence —	79
3.4.3.8	Verb phrase —	81
3.4.4	Overview results ICE Great Britain —	83
3.5	Null subjects in ICE Great Britain: First person contexts —	84
3.5.1	Overview: Subject tokens and linguistic factors first person —	84
3.5.2	Results first person —	85
3.6	Summary and discussion —	89
3.6.1	Discussion full data set —	89
3.6.2	Discussion first person —	92
3.6.3	Conclusion —	94

Part B: Null subjects in Asian Englishes

4	Asian languages and varieties of English: Theory, description and comparison —	99
4.1	Null subjects in Asian languages —	100
4.1.1	Radical pro-drop —	100
4.1.2	Topic prominence —	104

4.1.3	Quantitative research on Chinese null subjects —	110
4.2	Language contact and varieties of English —	111
4.2.1	Language contact and contact languages —	112
4.2.2	Null subjects revisited: Acquisition and contact —	116
4.2.3	English in contact: World Englishes —	119
4.2.4	Null subjects in (other) varieties of English —	122
4.3	Asian varieties of English: Background and structure —	124
4.3.1	English in India —	124
4.3.2	English in Hong Kong —	130
4.3.3	English in Singapore —	134
4.4	Comparing varieties —	141
4.4.1	Asian Englishes —	141
4.4.2	The comparative variationist method —	145
5	Empirical comparison: Null subjects in Asian Englishes —	148
5.1	Linguistic factors in the Asian Englishes —	148
5.1.1	The envelope of variation —	149
5.1.2	Coreferential coordination —	151
5.1.3	Clause type —	152
5.1.4	Position —	155
5.1.5	Person —	155
5.1.6	Specific reference —	156
5.1.7	Switch reference —	158
5.1.8	Persistence —	158
5.1.9	Verb phrase —	159
5.2	Overview Asian Englishes: Descriptive statistics —	160
5.2.1	Different distribution of null subjects in the data sets —	160
5.2.2	Collocation: Asian Englishes —	164
5.2.3	Summary subject tokens and linguistic factors: Asian Englishes —	171
5.3	ICE India —	173
5.3.1	Indian English: Extra-linguistic factors —	173
5.3.2	Indian English: Logistic regression model —	176
5.3.3	Indian English: First person contexts —	185
5.3.4	Short summary: Null subjects in Indian English —	186
5.4	ICE Hong Kong —	186
5.4.1	Hong Kong English: Extra-linguistic factors —	186
5.4.2	Hong Kong English: Logistic regression model —	187
5.4.3	Hong Kong English: First person contexts —	195
5.4.4	Short summary: Null subjects in Hong Kong English —	197

5.5	Grammar of Spoken Singapore English Corpus —	197
5.5.1	Singapore English: Extra-linguistic factors —	197
5.5.2	Singapore English: Logistic regression model —	199
5.5.3	Singapore English: First person contexts —	208
5.5.4	Topic prominence in Singlish: Within and beyond the clause —	209
5.5.4.1	Chinese-style topics —	210
5.5.4.2	Topic chains —	211
5.5.4.3	Discourse orientation and ambiguity —	213
5.5.4.4	Repair —	216
5.5.4.5	Code-switching? —	218
5.5.5	Short summary: Null subjects in Singapore English —	219
5.6	Summary and Discussion: Null subjects in Asian Englishes —	220
5.6.1	Summary: Asian English null subjects —	221
5.6.2	Variety specific contexts and puzzles —	227
5.6.3	Substrate comparison: How Asian are Asian Englishes? —	228
5.6.4	Conclusion —	233
6	Discussion and Conclusion —	234
6.1	Superstrate comparison: How English are Asian Englishes? —	234
6.1.1	Comparative variationist measures: Results —	234
6.1.2	Comparison: First person pronouns —	239
6.1.3	Summary —	241
6.2	Variable rules, variationist typology and language contact —	243
6.2.1	Null subjects in non-NSL: Theoretical and methodological implications —	243
6.2.2	Language contact and linguistic nativisation —	247
6.3	Conclusion —	251
	References —	255
	Appendix A: Data —	275
	Appendix B: Additional statistics —	277
	Index —	289

List of Tables

Table 2.1:	Prince's taxonomy of given and new information — 19
Table 2.2:	Typology of null subject languages — 25
Table 2.3:	Frequency of initial ellipsis in AmE and BrE conversation; occurrences per mil. words — 35
Table 3.1:	Verb forms attracting null subjects ICE-GB — 63
Table 3.2:	Verb forms repelling null subjects ICE-GB — 64
Table 3.3:	Linguistic factors and token numbers ICE-GB — 65
Table 3.4:	Sources of linguistic factors — 66
Table 3.5:	Wald statistics of factor group significance ICE-GB — 68
Table 3.6:	Logistic regression model ICE-GB — 70
Table 3.7:	Estimated variance inflation factors ICE-GB — 71
Table 3.8:	Results for factor group coordination ICE-GB — 72
Table 3.9:	Results for factor group clause ICE-GB — 75
Table 3.10:	Results for factor group position ICE-GB — 75
Table 3.11:	Results for factor group person ICE-GB — 76
Table 3.12:	Results for factor group specific reference ICE-GB — 77
Table 3.13:	Results for factor group switch reference ICE-GB — 78
Table 3.14:	Results for factor group persistence ICE-GB — 80
Table 3.15:	Results for factor group verb type ICE-GB — 81
Table 3.16:	Comparison of log odds for verb type glm vs glmer — 83
Table 3.17:	Overview linguistic factors first person ICE-GB — 85
Table 3.18:	Wald statistics of factor group significance first person ICE-GB — 86
Table 3.19:	Logistic regression model first person ICE-GB — 87
Table 3.20:	Estimated variance inflation factors first person ICE-GB — 87
Table 3.21:	Comparison of factor group significance all vs first person ICE-GB — 87
Table 3.22:	Comparison of factor group significance first person contexts ICE-GB, US English, and Nfdl English — 93
Table 4.1:	Thomason and Kaufman's borrowing scale — 114
Table 4.2:	Developmental phases in Schneider's evolutionary cycle of New Englishes — 121
Table 4.3:	WAVE pronoun omission features in Asian varieties — 143
Table 5.1:	WAVE VP deletion features Asian Englishes — 150
Table 5.2:	Subject tokens all 4 corpora — 162
Table 5.3:	Verb forms attracting null subjects ICE-IN — 166
Table 5.4:	Verb forms attracting null subjects ICE-HK — 166
Table 5.5:	Verb forms attracting null subjects GSSEC — 166
Table 5.6:	Verb forms repelling null subjects ICE-IN — 169
Table 5.7:	Verb forms repelling null subjects ICE-HK — 169
Table 5.8:	Verb forms repelling null subjects GSSEC — 169
Table 5.9:	Linguistic factors and token numbers Asian Englishes — 172
Table 5.10:	L1 speakers ICE-IN — 175
Table 5.11:	Wald statistics of factor group significance ICE-IN — 176
Table 5.12:	Estimated variance inflation factors ICE-IN — 177
Table 5.13:	Logistic regression model ICE-IN — 177
Table 5.14:	Results for factor group coordination ICE-IN — 178

Table 5.15:	Results for factor group clause ICE-IN — 179
Table 5.16:	Results for factor group person ICE-IN — 179
Table 5.17:	Results for factor group switch reference ICE-IN — 180
Table 5.18:	Results for factor group persistence ICE-IN — 180
Table 5.19:	Results for factor group verb type ICE-IN — 181
Table 5.20:	Results for non-significant factor group position ICE-IN — 182
Table 5.21:	Results for non-significant factor group specific reference ICE-IN — 183
Table 5.22:	Logistic regression model first person ICE-IN — 185
Table 5.23:	Wald statistics of factor group significance ICE-HK — 188
Table 5.24:	Estimated variance inflation factors ICE-HK — 188
Table 5.25:	Logistic regression model ICE-HK — 189
Table 5.26:	Results for factor group coordination ICE-HK — 189
Table 5.27:	Results for factor group clause ICE-HK — 190
Table 5.28:	Results for factor group person ICE-HK — 191
Table 5.29:	Results for factor group switch reference ICE-HK — 191
Table 5.30:	Results for factor group persistence ICE-HK — 192
Table 5.31:	Results for factor group verb type ICE-HK — 193
Table 5.32:	Results for non-significant factor group position ICE-HK — 194
Table 5.33:	Results for non-significant factor group specific reference ICE-HK — 194
Table 5.34:	Logistic regression model first person ICE-HK — 196
Table 5.35:	Wald statistics of factor group significance GSSEC — 199
Table 5.36:	Estimated variance inflation factors GSSEC — 199
Table 5.37:	Logistic regression model GSSEC — 200
Table 5.38:	Results for factor group coordination GSSEC — 201
Table 5.39:	Results for factor group clause GSSEC — 202
Table 5.40:	Results for factor group position GSSEC — 204
Table 5.41:	Results for factor group specific reference GSSEC — 204
Table 5.42:	Results for factor group switch reference GSSEC — 205
Table 5.43:	Results for factor group persistence GSSEC — 205
Table 5.44:	Results for factor group verb type GSSEC — 206
Table 5.45:	Results for non-significant factor group person GSSEC — 206
Table 5.46:	Logistic regression model first person GSSEC — 209
Table 5.47:	Factor group significance and range: Asian Englishes — 224
Table 5.48:	Factor range and ranking of factor levels Asian Englishes - Chinese — 230
Table 6.1:	Comparison of factor group significance and range — 235
Table 6.2:	Comparison of Rbrul constraint hierarchy — 236
Table 6.3:	Comparison of factor group significance for first person contexts vs full data set — 239
Table A.1:	GSSEC files in ICE-Singapore — 275
Table B.1:	Mixed-effects model with speaker as random intercept ICE-GB — 277
Table B.2:	Mixed-effects model with verb token as random intercept ICE-GB — 277
Table B.3:	Analysis of Deviance model comparison including factor group turn length ICE-GB — 278
Table B.4:	Analysis of Deviance model comparison excluding factor group switch reference ICE-GB — 278
Table B.5:	Minimal adequate model ICE-GB - overview — 279

Table B.6:	Model evaluation by bootstrapping minimal adequate model ICE-GB (n = 142) — 279
Table B.7:	Analysis of Deviance model comparison with monofactorial model “and” ICE-GB — 279
Table B.8:	Logistic regression model including factor group VP complexity ICE-GB — 280
Table B.9:	Wald statistics of factor group significance including factor group VP complexity ICE-GB — 280
Table B.10:	Analysis of Deviance model comparison including factor group VP complexity ICE-GB — 281
Table B.11:	Logistic regression model including interaction term Person:VerbType ICE-GB — 281
Table B.12:	Analysis of Deviance model comparison including interaction term Person:VerbType ICE-GB — 282
Table B.13:	Minimal adequate model first person ICE-GB - overview — 282
Table B.14:	Model evaluation by bootstrapping minimal adequate model first person ICE-GB (n = 169) — 282
Table B.15:	High-frequency verb lemmas — 283
Table B.16:	Minimal adequate model ICE-IN — 283
Table B.17:	Analysis of Deviance model comparison minimal adequate model ICE-IN — 284
Table B.18:	Model evaluation by bootstrapping logistic regression model ICE-IN (n = 125) — 284
Table B.19:	Minimal adequate model ICE-HK — 284
Table B.20:	Analysis of Deviance model comparison minimal adequate model ICE-HK — 285
Table B.21:	Model evaluation by bootstrapping logistic regression model ICE-HK (n = 191) — 285
Table B.22:	Mixed-effects model with ethnicity as random intercept GSSEC — 285
Table B.23:	Minimal adequate model GSSEC — 286
Table B.24:	Analysis of Deviance model comparison minimal adequate model GSSEC — 287
Table B.25:	Model evaluation by bootstrapping logistic regression model GSSEC (n = 200) — 287

List of Figures

Figure 2.1:	Givón's topicality hierarchy — 20
Figure 3.1:	Percentage of clause types by person in ICE-GB conversations — 50
Figure 3.2:	Raw frequencies of high-frequency lexemes in ICE-GB conversations — 57
Figure 3.3:	Histogram of null subjects per file in ICE-GB conversations — 58
Figure 3.4:	Percentage of null subjects per speaker in ICE-GB conversations — 59
Figure 3.5:	Percentage null subjects by gender ICE-GB — 60
Figure 3.6:	Percentage null subjects by age group ICE-GB — 61
Figure 3.7:	Attraction and reliance scores of verbs to ICE-GB null subjects — 62
Figure 3.8:	Attraction and reliance scores of verbs in position AND1 ICE-GB — 73
Figure 3.9:	Attraction and reliance scores of verbs in position AND2 ICE-GB — 74
Figure 3.10:	Referential vs non-referential second person pronouns ICE-GB — 76
Figure 3.11:	Interaction graph person by reference ICE-GB — 78
Figure 3.12:	Interaction graph switch reference by coordination ICE-GB — 79
Figure 3.13:	Interaction graph switch reference by persistence ICE-GB — 80
Figure 3.14:	Interaction graph verb type by person ICE-GB — 82
Figure 3.15:	Predicted probabilities for factor levels ICE-GB — 83
Figure 3.16:	Predicted probabilities for factor levels first person ICE-GB — 88
Figure 4.1:	Meyerhoff's taxonomy of transfer — 146
Figure 5.1:	Comparison of absolute numbers of clause types by person — 154
Figure 5.2:	Referential vs non-referential pronouns — 157
Figure 5.3:	Comparison of null subjects per file — 160
Figure 5.4:	Amount of speakers with categorical overt subjects — 161
Figure 5.5:	Comparison percentage of null subjects per conversation — 162
Figure 5.6:	Comparison null subjects by coordination — 163
Figure 5.7:	Collocates of null subjects Asian varieties — 164
Figure 5.8:	Attraction and reliance scores of verbs in position AND1 Asian Englishes — 170
Figure 5.9:	Attraction and reliance scores of verbs in position AND2 Asian Englishes — 171
Figure 5.10:	Percentage null subjects by gender ICE-IN — 174
Figure 5.11:	Percentage null subjects by age group ICE-IN — 174
Figure 5.12:	Percentage null subjects by L1 ICE-IN — 175
Figure 5.13:	Interaction graph switch reference by persistence ICE-IN — 181
Figure 5.14:	Interaction graph verb type by person ICE-IN — 182
Figure 5.15:	Referential vs non-referential pronouns ICE-IN — 183
Figure 5.16:	Predicted probabilities for factor levels ICE-IN — 184
Figure 5.17:	Percentage null subjects by gender ICE-HK — 187
Figure 5.18:	Interaction graph switch reference by persistence ICE-HK — 192
Figure 5.19:	Interaction graph verb type by person ICE-HK — 193
Figure 5.20:	Referential vs non-referential <i>it is</i> ICE-HK — 194
Figure 5.21:	Predicted probabilities for factor levels ICE-HK — 195
Figure 5.22:	Percentage null subjects by ethnic group GSSEC — 198
Figure 5.23:	Interaction graph verb type by person GSSEC — 207
Figure 5.24:	Predicted probabilities for factor levels GSSEC — 208
Figure 5.25:	Comparison percentage null subjects by speaker — 219
Figure 5.26:	Factor group person in the Asian Englishes — 225

Figure 5.27: Factor group clause in the Asian Englishes — **228**

Figure 6.1: Comparison effects verb type full data set — **241**

Figure 6.2: Comparison effects verb type first person — **241**

Figure 6.3: Comparison predicted probabilities — **242**

List of Abbreviations

AAVE	African American Vernacular English
AGR	agreement
APiCS	Atlas of Pidgin and Creole Structures
CL	classifier
CSC	complex stative construction
CRS	currently relevant state
DEM	demonstrative
DUR	durative
EFL	English as a foreign language
EPP	extended projection principle
ESL	English as a second language
eWAVE	electronic World Atlas of Variation in English
EXP	experiential
GSSEC	Grammar of Spoken Singapore English Corpus
HK	Hong Kong
ICE	International Corpus of English
IN	India
INDF	indefinite
IQR	Interquartile Range
LE	aspect particle (Mandarin Chinese)
LRT	Likelihood Ratio Test
MUH	morphological uniformity hypothesis
NfdIE	Newfoundland English
NP	Noun Phrase
NSL	null subject language
NSP	null subject parameter
PART	particle
PFV	perfective
POSS	possessive
Q	question particle
SAIE	South African Indian English
SBCSAE	Santa Barbara Corpus of Spoken American English
SG	Singapore
SGEM	Speak Good English Movement
SLA	second language acquisition
SP	subject-prominent
SPP	subject personal pronoun
TAM	tense-aspect-mood
TL	target language
TP	topic-prominent
VIF	variance inflation factor
VP	Verb Phrase
WALS	World Atlas of Language Structures

1 Introduction

1.1 Null subjects

Few grammatical structures have engaged as much attention from linguists of all theoretical backgrounds as the variable occurrence of subject pronouns in different languages and language types. The crucial role of grammatical subjects in the syntax of European languages makes their realisation a widely employed measure for typological classification of the syntactic configuration of a language. English as the prototype case of non-null subject languages has had a central role in these debates, especially in the productive era of the Principles and Parameters framework from the late 1970s onwards (see e.g. Chomsky 1981, Rizzi 1982, Rizzi 1986 for formative accounts). On the other hand, the Asian languages in contact with English in India, Hong Kong, and Singapore represent a contrary syntactic configuration, and remould the regional varieties of English accordingly. This tension, together with the quantitative nature of the syntactic variable, make null subjects in varieties of English, specifically the Asian Englishes, a fruitful field of investigation for researchers interested in the interface of typology, language-internal variation, and variationist sociolinguistics. Although the linguistic discourse on null subjects has crucially been shaped by formalist discussion, to account for language-internal microvariation in subject realisation the present study will present an empirical quantitative perspective on null subjects in spoken Englishes.

Null subjects are commonly associated with rich morphological systems. This is true for a number of languages that regularly use null subjects with tensed verb forms, e.g. Spanish, Italian, or Greek. The respective expressions are ungrammatical in languages with impoverished verbal morphology like French and English (1.1).

- (1.1) Ø Parla italiano.
 Ø Habla español.
 Ø Mila ellinika.
 *Ø Parle français.
 *Ø Speaks English.

In inflectional null subject languages (NSL), unambiguous morphological marking of the verb phrase indicating person and number of the subject carries the function of referent identification and makes its overt lexical expression redundant. In fact, overt subject pronouns can be considered the marked case in

these *canonical* NSLs, serving communicative functions like expressing emphasis or contrastive focus. Another type of NSL is represented by highly analytic languages like Chinese or Japanese, where the licensing of null pronouns is provided by the discourse context rather than by morphological marking. This language type is referred to as *radical* or *discourse pro-drop* language (Huang 1984). The apparent dichotomy between inflectional and analytic NSLs is addressed by the Morphological Uniformity Hypothesis, which predicts the absence of null subjects from languages with reduced or mixed morphological paradigms (Jaeggli and Safir 1989a). Compared to languages with obligatory overt subject pronouns, both canonical and radical NSLs are very common from a cross-linguistic perspective (Dryer 2013).

The possible contribution of quantitative approaches to the study of languages with optional subject pronouns is evident: the amount of null subjects compared to overt pronominal subjects varies both between and within languages classified as NSLs of some sort, ranging from 11% null subjects in Finnish, to 72% in Italian (Torres Cacoullos and Travis 2014: 21), or varying between dialects and registers of the same language, such as rates between 50% to 75% null subjects in regional varieties of Spanish (Torres Cacoullos and Travis 2015: 82).

English, on the other hand, usually serves as a model of non-null subject languages, representing the “Standard Average European” obligatory requirement for overt subject pronouns (Haspelmath 2001). However, despite this prototypical status examples like *(do you) know what I mean?*, *(I) don’t think so*, or *(it/this) sounds good to me* are acceptable and commonly encountered in informal speech. Some dialects of English are described as having a greater predisposition towards pronoun omission, e.g. Newfoundland English (Clarke 2004, Wagner 2012); additionally, null subjects are attested more frequently in specific registers like diaries (Haegeman and Ihsane 2001). The amount and conditioning of null subjects in Standard spoken English, British English specifically, remains largely in the dark as far as quantitative evidence is concerned. One notable exception is the preliminary account of null subjects in British and American English in Biber et al. (1999: 1105–1106). Their broad overview of the absolute numbers of null subjects found in reference corpora of British and American English points towards measurable differences in the frequency of null subjects between the two Standard varieties, with British English showing higher frequencies. This calls for further investigation, especially given the role of British English as historical input variety for many World Englishes, Asian Englishes specifically. The first part of the present study is hence devoted to the analysis of null subjects in spoken Standard British English.

The generative position is that such “non-canonical” null subjects “have special properties that distinguish them from the canonical null subjects” in lan-

guages like Spanish (Roberts and Holmberg 2010: 5). However, based on their analysis of spoken US English Travis and Lindstrom (2016: 104) state that concerning null subjects

English is not “special” at all, and the constraints in operation [...] are parallel to those widely described for null subject languages, suggesting a need for further exploration into the widely accepted distinction between “null” and “non-null” subject languages.

To test these conflicting hypotheses is one major motivation for this study, which joins the growing body of research empirically evaluating predictions made by theoretical approaches. This increasingly incorporates languages that do not neatly fit the prototypical categories but represent intermediate or mixed forms, such as the partial NSL Finnish (Helasvuo and Kyröläinen 2016), or creole languages that deviate from their superstrates with regard to subject expression (Meyerhoff 2000, Nicolis 2008).

A similar case is presented by the Asian contact varieties of English investigated here, whose NSL substrates gnaw away at the categorical demand of the Standard English superstrate for overt subject expression. Building on the results achieved for British English, the second part of the study analyses the effects of language contact on variable subject realisation in Indian English, Hong Kong English and Singapore English.

1.2 Research context and method

The focus on contact varieties places this study in one of the most vibrant fields of English linguistics, the research on the so-called *New Englishes* (Platt et al. 1984), or *World Englishes* (Kachru 1982). These post-colonial contact varieties of English have drawn considerable interest from linguists for several decades now. Their dynamic developments provide researchers with a real-time window to language evolution, granting insights into central issues of linguistic theorising like the range or limits of language change in contact situations, the flexibility or resilience of typologically marked and unmarked structures, the role of more general cognitive processes in language learning, and the implications for theories of language acquisition, language formation and the universal core of human grammar (Mesthrie and Bhatt 2008, see also various contributions in Filppula et al. 2017). Moreover, the individual stories of different varieties feature varying socio-historical settings, both historically in the founding phase, diachronically through the formation phase, and synchronically regarding the current status of a variety. Tracing these developments offers perspectives on the influence of indi-

vidual and societal identity formation and negotiation, which is often at odds with prescriptive governance of language policies, commonly guided by political and economic considerations. This interplay of forces results in widely varying linguistic outcomes (Schneider 2007).

Research on the linguistic systems of World Englishes is fiercely empirical, with a focus on corpus-based analysis of authentic language data, and a strong grounding in sound quantitative methodology. Variationist methods play an increasingly important role in these analyses. Quantitative variation of linguistic variables is a primary investigative field of *variationist sociolinguistics*, which has developed sophisticated statistical methodology to translate the perceived chaotic variation into insightful categories guiding the *orderly heterogeneity* within linguistic systems (Weinreich et al. 1968). The method of choice is *variable rule analysis*, conducted via binary logistic regression, which is also the statistical method employed in the present study (Tagliamonte 2006). The search for determinants of variation includes extra-linguistic factors, such as the classical sociolinguistic variables gender, age, and socio-economic class, and structural linguistic factors, which have repeatedly been shown to be more informative predictors of subject realisation and thus form the backbone of the present analysis. The combination of variationist methodology with a typological angle on the variables and factors selected for analysis seems a natural outcome of the World Englishes enterprise, whose field of study regularly features situations of contact between typologically diverging languages. Similar approaches are taken by e.g. Brunner (2014, 2017), Green and Meyer (2014), Siemund (2013, 2016), Siemund and Davydova (2017), and the contributions in Lim and Gisborne (2011).

Adopting a comparative perspective on the Asian Englishes by contrasting these underlying determinants of subject realisation allows for more qualified statements on the degree and amount of contact effects in each individual variety. The aim of the comparative variationist method is to evaluate effects of language contact by measuring structural convergence, visible in similarities and deviances between the linguistic constraints at work in different varieties (Tagliamonte 2008, Meyerhoff 2009). It is thus an ideal framework for this kind of study (see also Torres Cacoullos and Travis' forthcoming sketch of a "variationist typology").

The three Asian varieties investigated here have drawn considerable interest from linguists, and are thoroughly described structurally. While null subjects are well known as a feature of Asian Englishes in descriptive accounts (e.g. Pingali 2009 on India, Setter et al. 2010 on Hong Kong, Lim 2004 on Singapore, the respective chapters in Kortmann and Lunkenheimer 2012), quantitative investigation systematically integrating aspects of language contact, as is done e.g. for null copula by Ho and Platt (1993) on Singapore English, or as a comparative study by Sharma and Rickford (2009) and Sharma (2009) is lacking so far for null subjects

(but see Mesthrie and Bhatt 2008: 167–171 sketching a parametric approach to null subjects in World Englishes, or Sato and Kim 2012 for a recent discussion of the Morphological Uniformity Hypothesis and its validity for contact varieties like Colloquial Singapore English). The present study sets out to fill this gap and contribute to a deeper understanding of variable grammatical systems in contact varieties. The following section provides an overview of the databases used for this investigation.

1.3 Database

The International Corpus of English Project (ICE) is an ongoing international enterprise collecting and publishing data from varieties of English worldwide. All countries included “count as ‘English-speaking’ in some sense, but in fact are as culturally and linguistically diverse as Great Britain, Australia, India, Singapore, Hong Kong, Nigeria, or Fiji” (Nelson et al. 2002: xi). This excludes countries where English is exclusively used as a foreign language, such as most European countries, but includes a wide range of local communities of English speakers, including countries from Kachru’s Inner and Outer circles of World Englishes (Kachru 1985), and spanning the whole range of Schneider’s developmental phases in the evolutionary cycle of New Englishes (Schneider 2007, see section 4.2.3). The speakers included in any ICE corpus must be at least 18 years of age, and are required to be natives of the respective country. Within these countries, they need to have completed English-medium education at least until the completion of secondary school (Greenbaum 1996: 6). While data collection is ongoing, and many components of ICE are rather recent additions, the ICE components investigated here, ICE-GB, ICE-India, ICE-Hong Kong and ICE-Singapore, were collected in a similar timeframe (1990s and early 2000s).

Each ICE corpus contains one million words, a number that seems feeble with the recent advent of web-based mega corpora like the Corpus of Global Web-Based English (GloWbE, Davies 2013), but is sufficient for the investigation of most high- and medium frequency structural features. All ICE corpora follow a shared design, incorporating different registers and modes. They contain both spoken and written language of different registers such as monologues, dialogues, and edited as well as unedited writing (for an overview, see Greenbaum 1996). Each of the 500 texts included contains about 2,000 words, resulting in 600,000 words of spoken and 400,000 words of written English for each variety. According to Nelson, the possibility of comparison between global varieties of English is one key motivation for the global cooperative project:

While each component corpus can exist independently as a valuable resource for investigation into individual national or regional varieties, the value of the corpora is enhanced by their comparability with each other. (Nelson 1996: 27)

Genre variation is crucial for many morphosyntactic phenomena, and has repeatedly been shown to substantially influence the amount of null subjects (e.g. Li et al. 2012 on Chinese, Travis 2007 on Spanish, and Teddiman 2011 on English), but it is not resolved conclusively to which degree the underlying constraints on null subjects vary between genres (Travis and Lindstrom 2016). To secure comparability both between the varieties within the present study, and with previous studies on English, spoken data from the “direct conversations” register is used exclusively (ICE file category S1A, containing approximately 180,000 words per variety). Telephone conversations are also part of the “private dialogues” sub-corpus in ICE, but are excluded from the present study because they lack the shared environment and physical cues of face-to-face conversations. Not only do previous corpus studies on null subjects generally use the speech type “informal conversation” (see sections 2.2.2 and 2.3.2), it is also considered closest to naturally occurring speech (see e.g. Lange 2012: 4), and thus most likely to exhibit a non-standard feature like the omission of subject pronouns. This is also confirmed by Teddiman’s (2011) investigation of null subjects in different genres of ICE-GB, which finds the highest amounts of subject omission in private dialogues, compared to other spoken registers and the text type “correspondence”.

In the present study, the private conversation sub-corpus of ICE-GB is used to gain insight into the amount and types of null subjects in educated Standard English speech. The data contains a wide range of age groups. The subset of conversations analysed in detail varies in formality, ranging from doctor-patient exchanges and office hour conversations between students and supervisors, anecdotal narratives of elder relatives, to unforced conversations between friends and family members. In the Asian Englishes, the different societal backgrounds and functions of the three varieties naturally lead to slightly different kinds of data. Background information on the individual corpora is presented in the following.

ICE-India (ICE-IN) was released in 2002. The conversations were recorded between 1991 and 1997. The headers for the individual files provide information on age, gender, educational level, occupation and mother tongue and “other languages” of the speakers, who are largely trained as English language teachers (Shastri 2002). English in India serves as a lingua franca between speakers of different mother tongues (see also section 4.3.1). While some of the conversations appear genuinely informal and natural, others seem more stilted and resemble expository introductions of the speakers, or interview situations. Generally,

English is associated with higher formality, or even with elitism, as the following quote illustrates (1.2).

(1.2) If you speak English you are showing off <ICE-IN:010#117:A>

Similar issues are also addressed by Percillier in discussing his experiences with data collection in Malaysia and Indonesia: the formality of the speech situation influences the stylistic choice of speakers; on the other hand, depending on proficiency, not all lectal levels are available to all speakers within an ESL or EFL community (Percillier 2016: 192).

Data collection for the conversations used in ICE-Hong Kong (ICE-HK) was conducted between 1994 and 2001; the corpus was completed and made available for public use in 2006 (Nelson 2006). As for ICE-GB, information is provided on speaker age, gender, educational level, occupation, and linguistic background. Notable about ICE-HK is the substantive amount of non-Hong Kong speakers in the private dialogues, e.g. from the US, but also Japan, or the Philippines, indicated as speaker “Z” in the transcripts and not included in the word count. These speakers are obviously not part of the present analysis. The reason for their presence in the corpus becomes clear when the information on the mother tongue of the Hong Kong speakers is taken into consideration: Hong Kong is ethnically and linguistically rather homogeneous; all native Hong Kong speakers in the corpus are L1 speakers of Cantonese, there is no need for an inter-ethnic lingua franca like in India or Singapore, so English, especially for informal conversations, is only a necessary means in the presence of outsiders (see also section 4.3.2). A number of ICE-HK transcripts explicitly address difficulties in finding appropriate conversation partners for the data collection, and mention frequent code-switching in their usual everyday conversations, which is not considered desirable for the ICE data collection (1.3).

(1.3) You know in Hong Kong we often talk with people in a mixture of
English and Cantonese
A couple of my friends I mean they are Eurasian
For example I mention that but <,> because we speak half Chinese
half English so the conversation for this may not be so good <,>
<ICE-HK:033#20-22:A>

The Grammar of Spoken Singapore English Corpus (GSSEC) was given preference over ICE-Singapore (ICE-SG). Unlike ICE-SG, it contains information on age and ethnic group of the speakers, and provides recordings of the conversations together with the transcripts, which makes evaluation of their accuracy possible.

Several conversations from the GSSEC have been (partially) integrated into the private dialogue register of ICE-SG (see an overview in Table A.1 in Appendix A). The conversations were recorded from 1998 to 1999 by undergraduates from the National University of Singapore. The GSSEC files contain a total of approximately 60,000 words. They consist of rather relaxed and informal conversations between Singaporeans, mostly students and their friends and families, of all ethnic groups. The speakers are part of a generation that increasingly considers English, or more specifically, the colloquial variant Singlish, their native language (Lim and Foley 2004: 11–12). Several conversations directly discuss the role of Singlish as a marker of informality and local identity (1.4, see also section 4.3.3).

- (1.4) [W]hen we are delivering a lecture then I think
 it has to be of course of a certain standard not so much Singlish
 but the Singlish is used to illustrate [...] some jokes along the way
 to make the lecture more interesting
 and to also relate to the students who are [...] very localise in the
 Singaporean uh context <ICE-SG:071#70:A>

The introduction of the respective ICE components has provided a first glimpse at the different roles English plays in the post-colonial societies of India, Hong Kong and Singapore; this issue is addressed in more detail in chapter 4.

1.4 Scope and structure of the present study

The academic allure of the field of World Englishes is evident from the ever-growing number of dedicated journals, book series, edited volumes, handbooks, and monographs. So what can the present study contribute?

The objective of the present study is to paint a fine-grained picture of null subjects in spoken varieties of English in Asia, with the added bonus of providing the first such investigation for one of the Standard spoken varieties, British English. The detailed structural circumscription of null subjects in spoken Standard English constitutes a research gap that is only very recently being tackled by studies like Torres Cacoulios and Travis (2014), and Travis and Lindstrom (2016). However, both studies investigate Standard US English; so far no comparable analyses are available for British English, the shared historical input variety of the Asian Englishes investigated here, although differences between the two Standard varieties are likely, based on observations by Biber et al. (1999), and Haegeman and Ihsane (2001).

Contact varieties of English in South and Southeast Asia represent the interesting case of a blending of languages that are typically placed at the very extremes of the null subject vs. non-null subject scale, the local linguistic ecology of transplanted English is dominated by the discourse NSL type. The linguistic contexts investigated as possible determinants of subject expression thus combine results of cross-linguistic studies on null subjects (universal factors), descriptions and recent empirical investigations of English null subjects (English-specific factors), and insights from the study of language contact (contact factors). Given the role of this syntactic parameter in theory building, and the attested contact effects in the grammatical systems of the Asian Englishes, including subject pronoun realisation, providing the first in-depth investigation of this feature promises more profound insights into the processes and structural outcomes of radically different languages in contact.

The analysis is based on the transcribed conversational data from the respective ICE corpora. Though certainly interesting and worth further study, investigating the use of null subjects in different speech styles, such as narratives, or in written genres with differing formality exceeds the scope of the present study (see Teddiman 2011 for an overview on different genres in ICE-GB, and Tamaredo and Fanego 2016 for a first glimpse at other registers of ICE-IN and ICE-SG). The same is true for other ways of expressing reference and the management of information packaging in varieties of English influenced by topic-prominent substrates (see Lange 2012 on discourse patterns of Indian English, and Winkle 2015 providing a comparative perspective on information packaging in varieties of English). Furthermore, the influence of intonation on information structure and subject pronoun expression is worth its own dedicated investigation, and is thus not attempted here.

The aim of the study is twofold, and this is reflected in its structure, which is outlined in the following:

The first objective is to establish a baseline of what is the “acceptable” use of null subjects in spoken Standard British English, a benchmark that is not available so far. Chapter 2 provides the theoretical grounding for the quantitative investigation of null subjects in Standard English. Following a short overview of anaphoric reference and null subjects in theoretical accounts (section 2.1), tenets and methodological principles of the variationist paradigm are introduced (section 2.2). This includes a discussion of investigations of variable subject pronoun realisation in canonical NSLs, especially different (contact) varieties of Spanish. The goal is to identify candidates for universal structural constraints on null subjects. The status of English in NSL taxonomies is addressed, with specific focus on recent empirical studies inspired by variationist investigation of canonical

NSL, which contribute the “English-specific” constraints to the following analysis (section 2.3).

The analysis of null subjects in spoken British English, constituting the most comprehensive quantitative investigation of the phenomenon to date, is presented in chapter 3. A fine-grained description and illustration of the variable context (section 3.1) and the structural categories analysed (section 3.2) provides the opportunity for reconstructing and reproducing the analysis. Its results are presented as an overview of the frequency and types of null subjects found in the British data set (section 3.3), followed by a detailed discussion of the multivariate analysis conducted to evaluate the influence of the various structural factors (presented in section 3.4 and 3.5, and discussed in section 3.6). It can be confirmed that despite their comparatively rare occurrence, null subjects in British English follow regular structural constraints, and can meaningfully be analysed with variationist methods.

The second step of the investigation consists of a comparison of these results with findings from the Asian varieties. Chapter 4 provides the relevant background information on the structure of the substrate language type (discourse NSL), including a discussion of the few available studies within the variationist paradigm (section 4.1). This complements the discussion in chapter 2 regarding universal factors influencing subject pronoun realisation. Principles, determinants, and possible results of language contact are discussed, with a special focus on World Englishes and on null subjects in language contact (section 4.2). Given the importance of linguistic and sociohistorical ecology in the process of variety formation, this background is provided for each of the three varieties individually (sections 4.3.1, 4.3.2 and 4.3.3). These sections each include a short overview of relevant research on morphosyntactic features of the respective variety, with a closer look at accounts of null subjects specifically. The status of English in the different societies is evaluated in terms of Schneider’s dynamic model: Singapore English is further progressed in its nativisation status than the L2 varieties of India and Hong Kong, and this is reflected in a higher degree of structural deviance from the superstrate. As a methodological primer for the following comparative analysis, the comparative perspective on the varieties in question is adopted from a descriptive, and a methodological point of view in the spirit of the comparative variationist framework (section 4.4).

The goal of the comparative analysis conducted in chapter 5 is to relate the varying nativisation status and societal functions of different varieties of English in Asia to measurable differences in the realisation of a syntactic variable. The primary measures are quantitative differences, both in the attestation of null subjects in the speech community sampled in the respective corpora, and across different linguistic contexts. The analysis relies on contrastive diagnostics estab-

lished by the comparative variationist method. To ensure the comparability of results, this investigation follows the principles proposed in chapter 3 for British English in methodology and structure. Observations on divergent patterns in the variable context, the linguistic factors investigated, and the distribution of null subjects in the different data sets are addressed and illustrated (sections 5.1 and 5.2). The logistic regression models for the three Asian Englishes are first discussed separately (sections 5.3, 5.4 and 5.5), followed by an explicit comparison of the Asian varieties, and an evaluation of substrate influence (section 5.6). It is confirmed that the two L2 varieties Indian and Hong Kong English show more similarities with each other than with the more indigenised L1 variety Singapore English.

The study concludes with the comparative discussion of the results achieved on the four varieties and their implications in chapter 6. Additional material and complementary statistical information are provided in the Appendix.

Part A: **Null subjects in British English**

2 Null subjects: Theoretical, methodological and descriptive foundations

Subject personal pronoun alternation is a classic sociolinguistic variable, subject to multiple constraints. Indeed, considering the number of studies that have been carried out, [...] subject personal pronoun variation seems to have become something of a showcase variable in variationist sociolinguistics. (Bayley et al. 2012: 49–50)

While Bayley et al.'s statement holds true for Spanish sociolinguistics, and the syntactic phenomenon of null subjects has received scholarly attention from various perspectives, variationist research on English null subjects is scarce. The present chapter serves to establish the theoretical background for such an investigation, which is presented in the empirical analysis of variable subject pronoun realisation in spoken British English in chapter 3. Section 2.1 provides a general overview of the phenomenon from a cross-linguistic perspective. A short sketch of the forms and functions of pro-forms (section 2.1.1) is followed by an account of their role in different analytical approaches, both functional (section 2.1.2) and formal (section 2.1.3). While formal analyses approach the issue of null subjects quite differently both conceptually and methodologically, their influence on the debate has been decisive, and several empirically testable hypotheses can be derived from these theoretical treatments.

Section 2.2 lays the methodological grounds by introducing the research paradigm underlying the present study. The principles of variationist analysis are discussed (section 2.2.1), followed by examples of research on null subjects within this paradigm in canonical null subject languages and cross-linguistic contexts (section 2.2.2).

The role of subject pronouns in English is discussed in section 2.3.1, followed by an overview of quantitative investigation conducted on null subjects in native English (section 2.3.2).

2.1 Pro-forms and pro-drop

This section provides the foundation for the analytic factors used in the studies discussed in sections 2.2 and 2.3, and consequently in the quantitative analysis of variable subject realisation in chapters 3 and 5. As a general introduction to the concrete morphosyntactic phenomenon under investigation, different kinds of pro-forms as a grammatical and communicative phenomenon are described

(section 2.1.1), followed by an account of their role for information structure (section 2.1.2). The formalist perspective on the expression and omission of subject pronouns is addressed in section 2.1.3, introducing the distinction between *pro-drop* and *non-pro-drop*, or *null subject* and *non-null subject* languages.

2.1.1 Pro-forms

First and foremost, language is a communicative enterprise between speaker and hearer. Be it in dialogic and conversational, or narrative form, two major demands on successful communication are to grant informativity of utterances and ensure the interpretability of discourse. To achieve this goal, speakers need to structure their output beyond the sentence level with regard to both cohesion and coherence (Halliday and Hasan 1976).

Cohesion is defined as the “interpretation of some element in the discourse [being] dependent on that of another” (Halliday and Hasan 1976: 4), and supports the construction and configuration of meaning on the text level, i.e. coherence. Cohesion can take the form of *reference*, *substitution*, *ellipsis*, *conjunction*, also subsumed under *grammatical cohesion*, and *lexical cohesion* (Halliday and Hasan 1976: 6). Especially grammatical cohesion is relevant for the investigation here, since languages show systematic differences regarding their usage of linguistic means to achieve cohesion. As cohesive devices, *pro-forms* have the textual function of acting as covert links between sentences and clauses (Wales 1996: 4). According to Quirk et al. (1985), pro-forms are

words and word-sequences which are essentially devices of recapitulation or anticipating the content of a neighbouring expression, often with the effect of reducing grammatical complexity. (Quirk et al. 1985: 76)

Cataphora, the anticipation of reference, is more common for written rather than spoken language, which is the object of the present investigation, thus the following discussion is largely restricted to *anaphora* (the recapitulation of reference), as in (2.1).

- (2.1) Did you paint the portrait in the studio?
Yes, I did it there. (Chu 1998: 282)

The use of pro-forms reduces linguistic material and thereby condenses information; their use is thus more common in registers with lower informational density and shared situational context (Biber et al. 1999: 235). Semantically light

verbs like *do* (see also 2.1) and nouns like *thing*, *person*, etc. which are referentially vague, can act as pro-forms; the most common pro-forms, however, are *pro-nouns*, such as *I* and *it*, in 2.1 (Wales 1996: 5). The functional categories of pronouns introduced here are *personal* or *referential* pronouns, *generic* pronouns, and *syntactic* pronouns.

Personal pronouns are the “most prototypical” pronouns (Wales 1996: 1). In English, personal pronouns formally distinguish several categories, namely person, and partly number, gender and case (generative “phi-features”, Huang 2014: 174). They typically refer to specific language-external referents (Biber et al. 1999: 328). One crucial difference within this category is between first and second vs third person pronouns, as speech participants vs third party referents. Furthermore, the interpersonal first and second person pronouns are used primarily in their situational context, with shifting reference throughout interactive speech situations (Wales 1996: 3, see also the shift in form, with continuous reference, from *you* to *I* in 2.1).

Substitutional use is the most central function of third person pronouns. Typically, the occurrence of pronouns, as opposed to the usage of a full noun phrase, is accounted for by avoiding “redundancy of expression” (Quirk et al. 1985: 82). Third person pronouns can take the form of personal pronouns *he / she / they*, but also inanimate *it*, or demonstratives *this / these* and *that / those*. While first and second person pronouns are usually exophoric, i.e. referring deictically to speech-external entities, third person pronouns can denote both exophoric and endophoric, or textual, reference (Wales 1996: 44). According to Halliday and Hasan, the first instance of *it* in (2.2) refers to the physical act of curtsying, constituting extended exophoric reference. The second instance, referring to the whole of the Queen’s statement, or a *fact* rather than a *thing*, illustrates textual reference, although the difference is admittedly gradual (Halliday and Hasan 1976: 52).

- (2.2) [The Queen said:] ‘Curtsey while you’re thinking what to say. It saves time.’
 Alice wondered a little at this, but she was way too much in awe of the Queen to disbelieve it. (Halliday and Hasan 1976: 52)

In terms of their relative frequencies, the main anaphoric elements in conversational speech are personal pronouns, followed by definite NPs, while demonstrative pronouns are comparatively rare compared to other registers (Biber et al. 1999: 237). Especially in dialogues, the cohesive function of pronouns can also be fulfilled by a *null variable* (alternatively referred to as *gap*, *ellipsis* or *zero*, Hall-

iday and Hasan 1976: 142). According to Chomsky (1981: 65), who terms this the *avoid pronoun principle*,

[this] might be regarded as a subcase of a conversational principle of not saying more than is required, or might be related to a principle of deletion-up-to-recoverability, but there is some reason to believe that it functions as a principle of grammar. (Chomsky 1981: 65)

These determinants of recoverability are addressed in sections 2.1.2 and 2.1.3.

A less specific meaning is carried by the impersonal or generic usage of personal pronouns; this generalised exophoric reference to “people in general” can be made by *one*, or, more informally, plural pronouns of all persons (2.3).

- (2.3) Science tells us that the earth goes round the sun.
You can never tell what will happen.
They say it’s going to snow today. (Quirk et al. 1985: 353–354)

Purely syntactic pronouns (or: *expletives*, *pleonastic*, or *dummy pronouns*) constitute a cross-linguistic rarity (Newmeyer 2005). English uses *it* and *there* as expletive pronouns. Camacho’s (2013: 16–17) “taxonomy of expletives” provides four sub-categories (2.4). While the syntactic pronouns in a. (*weather ‘it’*) and b. (*extraposition* expletive), have “some kind of referential capability”, c. and d., *raising* and *existential* expletive are “purely expletive” (Camacho 2013: 16).

- (2.4) a. Here it always rains.
 b. It surprised me [that Santos won].
 c. It seems that she knows the truth.
 d. There are some visitors at the door. (Camacho 2013: 16)

Pleonastic pronouns are semantically empty and fulfil an exclusively syntactic function, whereas the use of referential pronouns is strongly determined by the *information status* of their referent. The following section introduces basic terms and concepts used to analyse these relations at the interface of syntax, semantics and pragmatics.

2.1.2 Pronouns and information structure

Information structure (Halliday 1967) or *information packaging* (Chafe 1976) refer to the way linguistic structure is employed to mark differences in the semantic categories of *givenness* and *newness*. Sentences are typically built according to

this two-fold division by placing given before new information (Halliday 1967: 205); corresponding oppositions are *theme* or *topic*, and *rheme*, *comment* or *focus* (for a more comprehensive account, see Lambrecht 1994: chapter 1). Structurally, these semantic notions are commonly associated with *subject* and *predicate*.

The use of referential pronouns is crucially determined by information packaging. Givón (1995) identifies iconic principles underlying the choice of referring expressions as processing signals. Following the *quantity principle*, more coding material is devoted to information less predictable for the hearer (Givón 1995: 49–50). The referential form is thus chosen based on the speaker’s hypothesis on the cognitive status of the referent in the hearer’s mind (see also Chafe 1976: 30). Lambrecht presents a twofold classification of this cognitive status of referents as more or less *identifiable* and *activated*. High degrees of identifiability and activation are achieved by three major means:

- a. introduction into the linguistic context;
- b. physical presence in the extra-linguistic context;
- c. retrievable from long-term memory as part of encyclopaedic knowledge. (Lambrecht 1994: 79–80)

For communicatively successful reference to language external entities, especially when absent from the immediate context, those referents usually have to be introduced into the linguistic context by full NPs in various syntactic functions. According to Ariel (1994), the *accessibility hierarchy of referring expressions* governs the linguistic codification of referents by universal psychological principles of cognitive accessibility. The decisive factor for the choice between nominal and pronominal expression is the “assumed familiarity” of the respective referent, as modelled in Prince’s *taxonomy of given – new information*, which classifies referents according to two cross-cutting distinctions, based on the dimensions “discourse” and “hearer” (Table 2.1, adapted from Prince 1992: 301). (2.5) is provided by Ward and Birner (2004) as an illustration of this taxonomy.

Table 2.1: Prince’s taxonomy of given and new information

	Hearer-old	Hearer-new
Discourse-old	Evoked: <i>it</i>	(non-occurring)
Discourse-new	Unused: <i>the president</i>	Brand-new: <i>a speech</i>

- (2.5) The president gave a speech today, and in it he offered a new tax plan.
(Ward and Birner 2004: 156)

In English, the function of personal pronouns (vs full NPs) is not unlike the use of definite articles with NPs, and only possible for “hearer-old” referents, as opposed to indefinite articles, or bare NPs for “brand-new” referents (see e.g. Wales 1996: 11). Further distinctions are drawn on the basis of the different discourse status: While the definite NP *the president* is discourse-new, it can be identified via Lambrecht’s “encyclopaedic knowledge” and is thus treated as “hearer-old”. The indefinite NP *a speech* represents the least familiar entity. Consequently, as the direct object it is located at the right periphery of the clause after the more familiar referent, and is further elaborated in the second part of the sentence. Once it is evoked, i.e. both hearer-old and discourse-old, reference via the anaphoric pronoun *it* is possible. Givón’s *topicality hierarchy* orders different structural means of reference according to their degree of givenness, or topicality (Figure 2.1, Givón 1983: 17).

high topicality

zero anaphora > weak pronouns > strong pronouns

right detachment > neutral order (+ definite) > left detachment

Y-movement > clefts > indefinite NPs

low topicality

Figure 2.1: Givón’s topicality hierarchy

Givón (1983: 18) states that the more continuous and predictable the referent, the less overt expression it needs to receive; highly salient antecedents possibly even license zero anaphora. The investigation here is mainly concerned with the denotation of referents via zero anaphora and pronouns, i.e. at the high-accessibility end of the topicality scale.

In a cross-linguistic study on referring expressions, including English and Chinese, Gundel et al. (1993: 275) model givenness as an implicational hierarchy of types of cognitive status, with decreasing givenness from left to right (Figure 2.2).

GIVEN

in focus > activated > familiar > uniquely identifiable > referential > type identifiable

NEW

Figure 2.2: Gundel et al.’s givenness hierarchy

With decreasing givenness, the scope of potential referents increases and thus becomes more ambiguous. Gundel et al. predict that the choice of anaphoric expressions is determined by the cognitive status of the referent: the more ambiguous the referent, the more explicit the referring expression, i.e. “in focus” elements can be realized by anaphors with the least phonological and morphological content, while “type identifiable” elements require full lexical NPs (Gundel et al. 1993: 282; this is also confirmed by Gundel et al. 2010). Gundel et al. (1993: 291–292) find that in each language of their cross-linguistic comparison, the two polar referential expressions are the most commonly employed choices. Overt pronouns in languages allowing for zero anaphora can thus be interpreted as the marked case, carrying a specific contrastive or emphatic function. However, pronoun realisation is usually variable, so accounting for null vs overt pronominal anaphora from such discourse-based perspectives bears the danger of circularity: a pronominal slot is empty, therefore its referent is deduced to be accessible enough for omission (Torres Cacoullos and Travis 2014: 20). In an attempt to more objectively operationalise givenness, the relative degree of cognitive accessibility of a referent is modelled as a function of antecedent-anaphor linear distance, or potentially interfering elements between antecedent and anaphoric expression within the text (e.g. by Ariel 1990, Givón 1983). High familiarity and activation status are natural concomitants of referential first and second person pronouns – both of them refer to physically present discourse participants. While different degrees of familiarity and activation are thus more meaningful for third person pronouns, distance to antecedent and possibly interfering referents are relevant for all pro-forms (see also Travis 2007).

The discourse status of its referent is also relevant for the syntactic role of subjects. The subject of a sentence is commonly defined as one of the two main constituents of a clause, e.g. by Biber et al. (1999: 122), as “the noun phrase that has a ‘doing’ or ‘being’ relationship with the verb in that sentence”. It can be identified via case marking in inflectional languages, and its role in syntactic processes like passivisation, reflexivisation, and government of the verb form (Kibrik 2001: 1413). Subjects are prototypically associated with the semantic role of *agent*. However, numerous examples of non-agentive subjects exist, e.g. instrument, recipient or affected subjects (Biber et al. 1999: 124), which requires the distinction between *syntactic* and *psychological subject*, visible e.g. in passive constructions like *The man was bitten by the dog*, where the syntactic subject represents the patient, while the psychological subject constitutes the post-verbal prepositional phrase. However, concerning its role in discourse, typically the subject coincides with the topic of the sentence, even in passive constructions (see e.g. Keenan 1976, Taboada and Wiesemann 2010). Following principles of information packaging, the principle of end-weight specifically, in SVO languages like

English the syntactic function of subject is usually filled by familiar referents, hence pronouns occupy the huge majority of subject slots (about 90%, vs 10% full NPs in ICE-GB; Biber et al. 1999: 237).

The choice of anaphoric expression is based on the language specific forms of anaphors, but is assumed to follow Givón's cross-linguistically valid quantity principle. In contrast, the role of non-referential or syntactic pronouns is less universal, and its connection with information structure is more indirect. Syntactically, the most explicit introduction of new referents in English consists of the existential construction with *there* BE in its presentational function (2.6, see e.g. Ward et al. 2002). In this construction, *there* acts as a pleonastic, purely syntactic subject, while the semantic correspondent to the subject function in the basic sentence pattern is displaced to post-verbal position.

- (2.6) There was a thing called a carvery which had a vast menu
<ICE-GB:021#146:C>

The existential construction enables the shift of the noun phrase to the post-verbal focus position, bypassing the restrictions imposed by the rather strict word order of English; syntactic structures with similar compensatory functions include raising constructions and extrapositions, which also require expletive pronouns as subjects in English (Ward and Birner 2004: 154, see 2.4 above). Existential presentational sentences like (2.6) introduce addressee-new entities, often followed by a relative clause further specifying the referent, and “promoting” them to topic status (Lambrecht 1988: 149). In accordance with Givón's topicality hierarchy, the introductory NP is usually indefinite. Indefiniteness is not an absolute constraint, though: definite NPs can also add (relatively) new information, either by serving as reminders, new subtypes, or specified by an identifying description (Ward et al. 2002: 1440). While these explicit presentational constructions are typical in narratives, in conversation new topics commonly evolve naturally as subtypes of previous topics, with less overt syntactic marking.

Many investigations on information packaging are based on the analysis of narratives, partly evoked in experimental settings to keep the content constant (e.g. Chafe's famous “Pear Stories”, Chafe 1980). Conversational data often behaves more unpredictably, so not all observations from studies on narratives are equally valid for this text type (see also Travis 2007, Travis and Lindstrom 2016). Typically, the constructions illustrated above are more prominent in monologic or written language, where direct feedback between producer and recipient of the message is precluded, while informal and dialogic speech additionally draws upon non-canonical structures, intonation and ellipses to mark contrast, focus and givenness (Hirschberg 2004).

From these considerations of information structure, it is possible to identify referential continuity as a likely enabling factor for zero anaphora. The following section is concerned with a complementary perspective on anaphora and the cross-linguistic distribution of null anaphora in the form of null subjects, introducing the syntactic typological parameter *pro-drop*.

2.1.3 Pro-drop

Generative linguistics has played a formative role in the debate on null subjects. Although the present study is not formalist in nature, generative treatments of the phenomenon offer a range of possible explanations for subject omission that can be operationalised as potential predictors in the empirical investigation presented in chapter 3 and chapter 5.

Formal accounts of grammar distinguish between language types based on syntactic parameters. Macro-parameters refer to a specific syntactic setting of a language that entails further structural properties (Chomsky 1981). Employing the metaphor of an “on/off-switch”, parameter setting is analysed as an essential process of L1 and L2 acquisition (see e.g. Hyams and Wexler 1993, Rizzi 2000a on L1; Phinney 1987, White 1985 on L2, see also section 4.2.2). One of these so-called *macro-parameters* is the *pro-drop*, or *Null Subject Parameter* (NSP). Since its postulation in Perlmutter (1971), Chomsky (1981), and Rizzi (1982), it has become one of the most thoroughly discussed theoretical issues in linguistics (see e.g. Jaeggli and Safir 1989b for a variety of accounts, Biberauer et al. 2010 for more recent approaches in the minimalist framework; Camacho 2013 for an overview).

In its original form, the NSP established the contrast between non-*pro-drop* and (canonical) *pro-drop* languages. According to the (extended) projection principle (EPP), all sentences require an overt grammatical subject (Chomsky 1982: 10). Depending on theoretical background, for empty subject slots three possibilities satisfy this principle: *little pro* as an empty category, zero as deleted pronoun, or inflection as pronominal (Camacho 2013: 60). The necessary condition for empty subjects in tensed clauses is the formal *licensing*, i.e. on the syntactic level, while *identification* is the sufficient condition, describing the semantic issue of recoverability of the referent of *little pro* (Rizzi 1986: 518–522).

In canonical null subject languages (NSLs), this is granted via inflectional marking or “rich agreement” (Chomsky 1981: 241; see also Rizzi 1986, Taraldsen 1978). The prototypical examples for this language type are Romance languages like Spanish or Italian, which specify both person and number via verbal morphology (2.7).

(2.7) a. — *vai* *al* *mare?*
 go-2.SG to-the sea
 ‘Are you going to the beach?’

b. — *non* *mangiamo* *carne.*
 not eat-1.PL meat
 ‘We don’t eat meat.’
 (D’Alessandro 2015: 202)

The complementary language type of a “well-behaved” non-pro-drop language is mainly represented by languages like English or French, which exhibit a diminished system of inflectional marking. This language type is predominantly found in Europe; in fact, Haspelmath (2001) categorises [– pro-drop] as a feature of “Standard Average European”. In contrast, null subjects are pervasive in some highly analytic languages like Chinese or Korean, the so-called *radical pro-drop* languages (Huang 1984, see also section 4.1). From this observation stems Jaeggli and Safir’s (1989b: 29–30) “uniformity of the paradigm hypothesis”, an attempt to delimit the amount of morphology considered “rich” enough for licensing zero (see also Camacho 2013: 27). According to this “morphological uniformity hypothesis” (MUH), languages with uniform morphological (non-)marking license null subjects, while languages with a mixed system do not. While much research has been devoted to the morphological uniformity hypothesis, the results remain inconclusive (see e.g. Neeleman and Szendrői 2007 for an overview, Sato and Kim 2012 and Sato 2014 on the morphological uniformity hypothesis confronted with evidence from Singapore English).

The NSP as a macroparameter is designed to entail other properties of the NSL as well, e.g. *that*-T effect, free inversion, thematic and non-thematic zero, and rich agreement (Rizzi 1982). This has been shown to be inadequate descriptively faced with cross-linguistic evidence, first by Gilligan (1987; criticism of the “classical version” of the NSP is also summarized in Huang 2000: 57–60). Gilligan confronts the predictions made by the postulated clusters of properties of NSLs with a cross-linguistic sample of more than 100 languages. One of the remaining strong correlations is the *referentiality generalization*, i.e. languages with thematic null subjects also have expletive null subjects. However, unlike Nicolis (2008) claims, this is not a universal correlation (one exception is Dominican Spanish, see Camacho 2013: 36).

Parameters are also assumed to play a role in L1 acquisition. The investigation of child language shows that null referential subjects are commonly attested in all language types up to age three (see Hyams 1986, Hyams and Wexler 1993).

However, Hyams' (1986) original assumption of a universal initial default setting to [+ pro-drop] has encountered criticism, most fundamentally by Valian (1990), who argues for a more nuanced approach incorporating variable input from adult language. Rizzi (2000b) analyses child null subjects as a root phenomenon, a structural restriction also shared by languages like Corsican and Brazilian Portuguese, but not attested universally.

Compared to the formerly dominant Principles and Parameters paradigm, more current approaches within the Minimalist Programme are rather concerned with microparametric, or language internal variation (Camacho 2013: 9). Their aim is to account for the attested conflicts of parametric predictions with cross-linguistic observations. Two interrelated presently relevant fields of investigation have been of specific interest in the last years: 1. Revising the EPP to account for different degrees of pro-drop, i.e. different frequency of null subjects, by broadening the classification matrix (e.g. Biberauer 2008, Biberauer et al. 2010), and 2. the investigation of contact varieties and creole languages, and typological shifts in these (e.g. Modesto 2008 on Brazilian Portuguese, Camacho 2013 on Dominican Spanish; Bayer 2013, Lisser et al. 2015, Meyerhoff 2000 and Nicolis 2008 on Creole languages).

The initial binary system of pro-drop vs non-pro-drop languages has been considerably broadened since its first incarnation, first incorporating isolating languages like Chinese as radical pro-drop, later adding subclasses of intermediate types, or, as Holmberg terms it, different null subject *parameters* (Holmberg 2010). Subtypes of NSL are usually distinguished on the basis of possible types of null subjects, especially referential status, or person. Biberauer (2010) provides a “typology of NSLs” (summarised in Table 2.2).

Table 2.2: Typology of null subject languages

Pro-drop type	Example	Null subjects	Verb agreement
Consistent / canonical	Italian	yes	rich inflection
Expletive / semi	German	yes, but only expletives	some inflection
Partial	Finnish	sometimes (restricted)	some inflection
Discourse	Chinese	yes (and null objects)	no inflection

Expletive subjects are rare cross-linguistically (Newmeyer 2005); while they are obligatory in languages like Norwegian, Danish, Swedish and English, even languages that possess them often express them variably in different contexts (e.g. Icelandic, also German and Yiddish impersonal expletives, which are obligatory only clause-initially, Camacho 2013: 44). Expletive- or semi-NSL grant the licens-

ing of *pro*, but not identification, therefore only permitting expletive, but not referential null subjects; this is also referred to as *topic-drop* languages (Liceras and Díaz 1999).

Partial NSLs like Finnish allow null subjects in “certain restricted contexts” (Roberts and Holmberg 2010: 11). They commonly omit generic pronouns (Holmberg et al. 2009). Omission of referential subjects usually depends on inflectional marking (e.g. different marking for Hebrew persons, or Irish synthetic vs analytic forms, Camacho 2013: 93). A general tendency for a split between the speech act pronouns first and second vs third person is observed by e.g. Cole (2010), Frascarelli (2007), and Sigurðsson (2011). Such a “split system” is also described by Meyerhoff (2000) for Bislama, which, in contrast to Finnish and Hebrew, favours third person null subjects (see also Camacho 2013: 29).

Furthermore, even within NSLs, the fulfilment of the sufficient condition for null subjects does not necessarily entail a zero variant. This results in vast differences of rates of null subject realisation. The amount of null subjects in spoken language ranges from 79% in Polish, 72% in Italian, 76% in Javanese, 53% in Cantonese, and 11% in Finnish, all of which are classified as “null subject languages of some sort” (Torres Cacoullos and Travis 2014: 22).

Language internal variation is also attested for canonical NSLs, e.g. by Torres Cacoullos and Travis (2015) for dialect and genre variation in Spanish, where rates for first person zero vary from approximately 50% to 75% (see also contributions in Wratil and Gallman 2011 on language-internal variation). While the overall rates of subject omission clearly differ between genres, it remains open whether the constraints on omission still remain constant across regional and stylistic varieties of a language, or are systematically different, as suggested by Roberts and Holmberg (2010: 5).

The accounts of null subjects in various languages above yield the following possible determinants of subject omission: referential status of the omitted pronoun, i.e. possibly a higher likelihood of omitted expletives; and the favouring of null subjects in contexts of licensing morphology, i.e. third person singular present tense inflected verb forms in English.

Together with observations of processes in L1 acquisition, where null subjects occur regularly for all language types, and the insights concerning partial pro-drop, it is obviously more appropriate to describe null subject vs non-null subject languages as a continuous scale, rather than a binary opposition. One part in positioning languages on this scale is the quantitative degree of variation, and the range of contexts in which null subjects are allowed in a given variety. Systematic quantitative investigation as provided in the present investigation can thus contribute to clarifying the status of languages and varieties on such scales,

and help to evaluate the status of proposed absolute and implicational universals.

2.2 Quantitative approaches

As discussed above, the generative account of subject pronoun omission, in the course of its general microparametric turn, has had to admit to the existence of mixed language types and language internal variation with regard to null subjects. The investigation of such variation from a quantitative perspective has been the domain of variationist studies since the 1960s. Fundamentals of variationist analysis are introduced in section 2.2.1, followed by a discussion of relevant research within this framework (section 2.2.2).

2.2.1 The variationist paradigm and grammatical variation

The search for order in the perceived chaos of language-internal variation has been an ongoing endeavour. Labov's groundbreaking work on Weinreich et al.'s (1968: 99–100) “structured heterogeneity” provides methodological grounds for a prolific community of researchers within the variationist paradigm.

Most functional categories can be expressed by two or more differing morphological or lexical items, this is referred to as different *variants* of a linguistic *variable*, e.g. *will* vs *shall* vs *going to* as expressions of FUTURE in English (Torres Cacoullos and Walker 2009). The assumption is that hypotheses on the (non-) occurrence of variants can be empirically tested by operationalizing and statistical analysis (Torres Cacoullos and Travis 2014: 24). One pillar of this analytical framework is the *principle of accountability* (Labov 1982: 30), as phrased by Tagliamonte:

This principle is fundamental to variation analysis; it dictates that all occurrences of the target variable must be taken into account, not simply one variant or another. [...] In other words, you must include all non-occurrences as well. (Tagliamonte 2006: 72)

Accountability thus means assessing the whole range of *variable contexts*. This is only possible via the delimitation of the *envelope of variation*, a demarcation of variable vs categorical, i.e. invariant contexts of a variable, through the detailed inspection of the data that is to be analysed (Tagliamonte 2006: 88). Like other syntactic variables, the investigation of null pronouns necessarily demands a function-based approach to identifying the variable context (Sankoff 1988: 151).

A crucial interest in variationist analysis is the identification of *variable rules* in the form of factors systematically influencing the choice of specific variants to different degrees (e.g. Labov 1969, Sankoff and Labov 1979):

Thus, in order to determine the status of a form, it is not its current *existence* in a variety which is decisive, nor even its rates of occurrence. [...] However, the *distribution*, i.e. precisely where it occurs in the language, as determined by the relative frequency of the feature across its different contexts of use, is taken to represent the underlying grammatical structure. (Tagliamonte 2008: 133, emphasis in the original)

The conditions of linguistic variants can consist of extra-linguistic factors such as speaker age, gender, class, education, etc., or different structural contexts, such as phonological or syntactic environments, as linguistic factors. Typically, the distribution of linguistic variants is determined by various factors, or constraints. To determine the variable grammars of different languages and language varieties means to identify such regular patterns, the *variable rules* (Bayley 2013: 118).

Within the envelope of variation, such probabilistic constraints can be identified via quantitative statistical methods (Travis and Lindstrom 2016: 104). The aim is to gain insight into the *variable grammar* “underlying the variable surface manifestations” (Poplack and Tagliamonte 2001: 94), even across “data sets of different sizes, with varying frequencies of forms” (Poplack and Tagliamonte 2001: 93). The classic variationist tool is multiple logistic regression, often in the form of a dedicated statistical package, GoldVarb, but increasingly conducted in open-source software R with additional software packages (this alternative and its advantages are discussed, amongst others, by Bayley 2013, Johnson 2009, Sankoff 2008, Tagliamonte 2006, Wagner 2012). This statistical procedure evaluates and weights the contribution of different predictors, or factor groups, and their individual factor levels in a multivariate analysis (see also section 3.3.5).

Multivariate analysis provides the analytical measures of statistical significance, *relative factor strength*, and *constraint hierarchy* (Tagliamonte 2008: 132–133). Especially for low-frequency phenomena, *constraint rankings*, i.e. the relative importance of individual factor levels, can serve as a crucial methodological tool that enables the comparison of variable grammars across varieties (Poplack and Tagliamonte 2001: 94).

The investigation in chapter 3 follows these methodological guidelines and presents a multivariate analysis of the variable occurrence of null and overt subject pronouns in British English. Null subjects as variation between pronoun and zero expression are considered a case of (near) synonymous variants of a linguistic variable and have thus inspired a wealth of research, especially in canonical null subject languages. Central insights from these studies are presented in the following section.

2.2.2 Variationist studies on null subjects

Most quantitative research on null subjects has been conducted on canonical null subject languages, but recently also on contact varieties thereof, and partial null subject languages. Especially Spanish and its varieties have inspired a rich selection of quantitative research from various perspectives on subject personal pronouns (SPP; for exhaustive research overviews, see e.g. Bayley 2013, Carvalho et al. 2015, Flores-Ferrán 2007). These studies can also serve as empirical models for studies on non-NSLs like English (as discussed in section 2.3.2). A selection of variationist analyses concerned with the establishment of variable rules for null subjects in canonical and partial NSLs is presented below, chosen by their relevance for the present investigation, especially with regard to their description and choice of linguistic factors, and their assessment of the cross-linguistic validity of their findings (Englebretson and Helasvuo 2014, Erker and Guy 2012, Travis 2007, Travis and Torres Cacoulios 2012). Especially interesting here are studies concerned with partial NSLs (Helasvuo and Kyröläinen 2016), and with the consequences of language contact on subject realisation (Nagy et al. 2011, Nagy 2015, Otheguy et al. 2007, Silva-Corvalán 1994, Torres Cacoulios and Travis 2010, 2015).

Research on Spanish null subjects has identified several recurring factors determining subject realisation. A common insight is the rather peripheral role of extra-linguistic factors like gender, age and social class (Flores-Ferrán 2007: 625). The most influential structural and discourse factors are switch reference, discourse connectedness, “i.e. continuity of subject, tense, and mood from the preceding tensed verb”, reference chains, subject person and number, semantic features of the verb, and tense-mood-aspect of the verb phrase (Bayley 2013: 17). Due to the syntactic structure of Spanish, investigation is exclusively concerned with referential personal pronouns; however, the definition of the variable context is not always made sufficiently transparent concerning the inclusion of e.g. non-specific personal pronouns, or set expressions like discourse markers (Flores-Ferrán 2007: 628).

Predictably, contexts that help referential identification of the null element favour subject omission, such as referential continuity, e.g. in the form of reference chains, and a high degree of discourse connectedness (e.g. Silva-Corvalán 1994). Conversely, dialects that exhibit high rates of neutralising morphological distinctions of verb forms, e.g. between different persons, also show higher rates of overt subject pronouns (e.g. Bayley 2013: 16–17, Cameron 1996).

Concerning subject person and number, a common observation is the higher preference of plural forms for null subjects (e.g. Bayley and Pease-Alvarez 1997). The most frequent overt subject pronoun is first person singular *yo* ‘I’ (Flores-Ferrán 2007), although different distributions have been observed for Caribbean and

South American Spanish, which favour overt second person singular and third person singular respectively (Otheguy and Zentella 2012). This calls into question universal, discourse-based explanations, and hints towards variety-specific patterns of subject realisation dependent on person.

For semantic features, among others Silva-Corvalán (1994) and Travis (2007) identify the tendency for overt subjects with “verbs of mental activities or states” and “communicative activity” (Flores-Ferrán 2007: 634). The correlation observed between overt first person subjects and *psychological* or *cognitive verbs* is explained by

the role they play in expressing epistemicity, as it is through use of the first person that speakers can weaken or strengthen their stance towards an utterance, by using expressions such as (*yo*) *creo* and (*yo*) *pienso* (‘I think’). (Travis 2007: 113)

Further evidence for cognitive verbs as a separate class is provided by the existence of a turn-position effect which is only valid for this category (Travis and Torres Cacoullos 2012: 741). Within the class of psychological verbs favouring overt subjects, Travis and Torres Cacoullos identify central prototypical constructions, most prominently *yo creo* (‘I think’) as a high-frequency blueprint for further analogical *yo* + COGNITIVE VERB realisations (Travis and Torres Cacoullos 2012: 743).

Frequency is also central to a recent approach by Erker and Guy (2012), which tackles the influence of the verb phrase on subject pronoun realisation from a different angle, “the role of lexical frequency in syntactic variability”. In their analysis of New York Spanish conversations, they model lexical frequency as a discrete rather than a continuous effect. Verb forms are divided into the classes *low* and *high frequency*, defined as “individual verb form constitut[ing] at least 1% of the corpus” (Erker and Guy 2012: 536), resulting in 13 high-frequency forms in Erker and Guy’s data set. New York Spanish is a contact variety and potentially exhibits higher pronoun realisation rates due to its contact with English (see also Otheguy et al. 2007). Overall, high frequency verbs show rather low pronoun realisation rates. Erker and Guy explain this by a conserving effect of high frequency forms that are processed as chunks and thus not affected by language contact to the same degree as low frequency forms (Erker and Guy 2012: 531–532). Furthermore, their findings point towards a “non-orthogonal influence” of lexical frequency, i.e. frequency is not an independent, direct effect, but interacts with other constraints: “each of the core constraints is weaker and less predictive of SPP use among infrequent forms and stronger and more predictive among frequent forms” (Erker and Guy 2012: 546). So far, the validity of their observation is inconclusive, with Bayley et al. (2013) unable to verify their claim. However,

Wilson (2014) provides further evidence supporting the conclusions of Erker and Guy.

Another structural constraint observed for Spanish subject pronouns consists of *priming*, or *persistence*, effects. Travis (2007) investigates genre effects on first person pronouns specifically. Despite widely differing subject realisation rates in conversations compared to narratives, the linguistic constraints are identical in both genres and remain constant for both New Mexican and Colombian Spanish (Travis 2007: 115). Furthermore, both genres show two distinct priming effects: overt subjects are primed by specific lexemes, whereas null subjects are structurally primed (Travis 2007: 104). The lesser duration of the priming effect in the conversational data is attributed to the lower continuity of TAM marking of verb phrases compared to narratives (Travis 2007: 131). In their investigation of Colombian Spanish, Travis and Torres Cacoullós (2012) find two distinct priming effects, both by previous coreferential, and immediately preceding token irrespective of reference (Travis and Torres Cacoullós 2012: 729–730). While overt coreferential priming for further overt realisation can be considered a lexical effect of sorts (Travis' 2005 “yo-yo effect”), the triggering effect of null realisation for further omission is “an apparently purely structural effect” (Travis and Torres Cacoullós 2012: 730).

While this set of factors is well-researched from a variationist perspective for the canonical null subject language Spanish, empirical research on other language types is rather lacking so far. A decidedly cross-linguistic, discourse-based angle is adopted for Englebretson and Helasvuo (2014, a special issue of the *Journal of Pragmatics*, vol.63). The contributions focus on the variable expression of first and second person pronouns in natural conversation, spanning a wide range of different language types, i.e. Peninsular Spanish, European Portuguese, English, Swedish, Javanese, and Finnish (Englebretson and Helasvuo 2014: 1–4). Like the present study, contributions in the volume are concerned with language-internal, micro-level variability, and aim to enlighten conditions for speakers' choice of the non-typical variant for the respective language. A common finding is that

subject expression or ellipsis is often locally contingent and construction specific, and maintained through frequency effects, grammaticalization, and local priming. (Englebretson and Helasvuo 2014: 3)

The variable realisation of first person singular pronouns in the partial NSL Finnish is investigated by Helasvuo and Kyröläinen (2016). This study is especially relevant here for its methodological currency and transparency, and investigates many of the same factors as the present study. Helasvuo and Kyröläinen analyse about 2,000 first person singular subjects in spoken Finnish with regards

to “constructional”, and “discourse and cognitive” factors (Helasvuo and Kyröläinen 2016: 268). Constructional factors are concerned with properties of the verb phrase, such as verb semantics, argument structure, tense-mood-aspect marking, and polarity (Helasvuo and Kyröläinen 2016: 269–272), while discourse and cognitive factors are operationalised as referential distance, syntactic complexity, lexical frequency of the verb, turn length and persistence (Helasvuo and Kyröläinen 2016: 274–277). Amongst the tested random effects speaker, verb, and conversation, random variation between individual speakers is most informative. The strongest structural effect on subject realisation is found for the discourse factor persistence, emphasising the importance of the sequential organisation of conversations. Other statistically significant effects favouring null subjects, although with a much smaller effect size, are found for low referential distance, short syntactic units, intransitive verb phrases, and motion verbs, especially compared to psychological verbs, which strongly favour overt subject pronouns (Helasvuo and Kyröläinen 2016: 282–284). Several structural factors known for canonical null subject languages can thus be confirmed for the partial null subject language Finnish.

An interesting testing ground for the cross-linguistic validity of constraints on null subjects are studies of language contact between NSLs and non-NSLs. While Otheguy et al. (2007), and Otheguy and Zentella (2012) encounter higher rates of overt subject pronouns for Spanish in New York, Silva-Corvalán (1994) finds no such evidence for Spanish in Los Angeles. In order to analyse the possible effect of English on New Mexico Spanish, Torres Cacoullós and Travis (2010) stress the importance of more detailed investigation of the constraints on null subjects in contact varieties, rather than merely reporting subject realisation rates. These are known to vary widely between both regional and stylistic varieties of Spanish, while structural constraints remain largely constant (Torres Cacoullós and Travis 2010: 189–190). In order to separate contact effects from cross-linguistic tendencies, a crucial point in the analysis is the identification of structural *conflict sites* between the languages in contact (see also Poplack and Meechan 1998). In the resulting grammatical system of the contact variety, these conflicts can manifest as differences in the direction or strength of effects (Poplack and Tagliamonte 2001: 101). While Torres Cacoullós and Travis find no clear differences in linguistic constraints between non-contact and contact varieties, a stunning insight of their study is the validity of priming even across language boundaries, i.e. English *I* favours overt *yo* in instances of code-switching (Torres Cacoullós and Travis 2010: 205).

Torres Cacoullós and Travis (2015) consider subject pronouns a “paradigmatic case for grammatical convergence in studies of US Spanish” and aim to provide an inter- vs intralinguistic comparison of constraints. They consider

cross-linguistic factors like discourse cohesion, activation, accessibility and interactional, pragmatic factors. Their aim is to provide “foundations for the study of pronoun expression in Spanish in contact with English” by detailed investigation of the structural constraints in both languages involved in the contact situation, an approach very similar to the present study (see also section 2.3 below, and section 4.4.2). A first major difference is found in the variable context for pronoun expression, which is much more limited in English than in Spanish.

Nagy et al. (2011) and Nagy (2015) provide a whole collection of cross-linguistic comparisons, investigating the null subject “heritage languages” Cantonese, Italian and Russian in contact with English in Toronto. Subject continuity is the most widely attested factor group. While no detailed data is available on the homeland varieties of the respective languages, Nagy et al. find no changes between different generations of speakers of the different heritage languages, indicating that subject realisation is not used as a sociolinguistic marker (Nagy et al. 2011: 143).

To sum up, the variationist investigation of null subjects in canonical null subject languages assumes a near-synonymous relation of overt and null pronominal subjects. Analysing determinants of variation reveals that extra-linguistic factors like age and gender are not as relevant as structural factors, while dialect and genre variation yield clear differences in the rate of subject realisation, but hardly affect the linguistic constraints.

The analysis of contact varieties shows that contact between typologically different languages possibly, but not necessarily influences subject pronoun realisation. These contact effects are only revealed in a detailed assessment of linguistic constraints in the languages in contact, as well as in the resulting contact variety. A crucial notion is that of structural “conflict sites”, which are measurable in the resulting differences in the constraint hierarchy between substrate, superstrate and contact language.

Unlike factors specific to the inflectional language type of Spanish, namely discourse connectedness and identifiability of (null) subject referent via morphological marking, and variety-specific patterns for different persons, factors determined by principles of language processing, i.e. switch reference as an operationalisation of mental accessibility of the referent, lexical frequency of the verb, and priming, can be assumed to be valid cross-linguistically and therefore for (varieties of) English as well. A further candidate universal is the strong association of overt first person pronouns with psychological verbs.

The increasing number of quantitative studies on non-canonical NSL, including contact varieties and partial NSLs like Finnish, has opened the way to the evaluation of the cross-linguistic validity of structural determinants established for canonical null subject languages, as attempted in the present study. Previous

descriptive (section 2.3.1) and variationist (section 2.3.2) approaches to null subjects in English are discussed in the next section.

2.3 Null subjects in L1 English

In generative frameworks, English is usually cited as *the* “well-behaved” non-null subject language; still, in specific circumstances null arguments can be found. The following section provides an assessment of the status of English as a non-NSL (section 2.3.1), followed by an account of the state of empirical research regarding null subjects in L1 English (section 2.3.2).

2.3.1 English as a non-null subject language?

Although English started out historically as a “German-style topic drop language”, its historical development largely eliminated the regular possibilities for null subjects (see e.g. Walkden 2013). Still, the fact that null subjects do occur in actual spoken English has triggered a number of studies in the generative paradigm and in descriptive accounts of Standard spoken English.

Despite their supposed ungrammatical status, null subjects have their place in descriptive grammars of English. Quirk et al. state that “of the clause elements other than the verb, the subject is the most important in that [...] it is the element that is most often present” (Quirk et al. 1985: 725). However, even in Standard present-day adult English, a certain amount of null subjects can be found. They are obligatory in non-finite subordinate clauses (generative “big PRO”), and quasi obligatory in imperatives (Biber et al. 1999: 219). Null subjects are optional but canonical in coreferential coordinations (2.8).

- (2.8) This gay guy who came into the pub completely fell in love with Ben
and Ø was like declaring his undying love. (Biber et al. 1999: 156)

Although they are not strictly grammatical, the omission of sentence initial subjects is frequently reported for informal spoken language (see e.g. Zwicky and Pullum 1983 on “informal style deletion”, 2.9).

- (2.9) Ø got in last night and Ø still haven’t unpacked
(Zwicky and Pullum 1983: 156)

Biber et al.'s Longman Grammar (1999) treats null subjects as a kind of ellipsis found especially in conversation. They state that this situational ellipsis “takes place when the subject of a declarative clause is omitted, normally at the start of a turn” (Biber et al. 1999: 1105). An analysis of the British and American English conversational sub-corpora of the Longman Spoken and Written English Corpus reveals a non-negligible amount of utterance-initial null subjects in both Standard varieties, provided as estimated occurrences per million words (Table 2.3 adapted from Biber et al. 1999: 1105–1106).

Table 2.3: Frequency of initial ellipsis in AmE and BrE conversation; occurrences per mil. words

Ellipsis of	AmE	BrE	overall
Subject	1,000	3,000	2,000
Subject + operator	< 1,000	1,000	1,000

Including subject + operator omission (see also 2.10 c.) raises the number of null subjects in Biber et al.'s corpora to more than 1,000 for American English and 4,000 for British English, respectively. Apparently, null subjects are more common in British than in American English. This is especially evident for specific lexemes. A case in point is provided by the verb form *depends*, which is found without overt subject in 30% of all cases in American English, and 60%, i.e. the majority of its occurrences, in British English (Biber et al. 1999: 1106).

Given these findings, Liu (2008: 275) criticises the “clear-cut” categorisation of English as non-NSL, especially with regard to the insufficient development of grammars and teaching material targeted towards speakers of null subject languages. He cites the following examples as evidence for null subjects in informal spoken English (2.10).

- (2.10) a. What's concubine?
 Ø Don't know, get a dictionary.
 b. Do you want me to go hire a video camera while I'm at it?
 Yeah, Ø be great.
 c. Ø Know what I mean?
 d. What happened to John?
 Ø Had an accident.
 e. Ø Sounds good to me.
 (Liu 2008: 280)

Liu argues that instances of null subjects in English like above possibly serve as positive evidence for learners with a null subject L1, hindering or delaying presumed parameter resetting in the acquisition process (Liu 2008: 285, see also section 4.2.2 below).

Given the importance of parameter setting for first language acquisition, combined with the relatively high frequency of null subjects in early child English, this domain is a productive field of study. Hyams (1994) and Rizzi (1994) both claim that null subjects in child English are restricted to the first position of non-*wh* root clauses and constitute instances of German-style topic-drop rather than Spanish pro-drop. Xiao (2002) goes as far as classifying early L1 English as topic- rather than subject-prominent, manifesting in frequent empty subjects and dislocation processes (see also section 4.1). These observations support generative characterisations of null subjects as a root phenomenon (but see Haegeman and Ihsane 1999: 126).

The specific register of diaries is another source of studies on English null subjects, analysed e.g. by Haegeman and Ihsane (1999, 2001), Teddiman and Newman (2007), and Weir (2012). While embedded null subjects are clearly less common than root null subjects in their diary data, Haegeman and Ihsane observe them in various types of subordinate clauses, such as declarative complement clauses with and without *that*, interrogative complement clauses, relative clauses and adjunct clauses (Haegeman and Ihsane 1999: 129, 2.11).

- (2.11) Next time Ø see him Ø must be alive.
(Haegeman and Ihsane 1999: 128)

Haegeman and Ihsane (2001: 333) report higher acceptability for such cases from speakers of British English compared to American English, which is in line with Biber et al.'s observations on the frequency of null subjects in either variety (see Table 2.3 above). Teddiman and Newman (2007) draw upon a corpus of online diaries to elucidate the distribution of null subjects in this rather informal written variety. They confirm findings by Nariyama (2004) on higher omission rates for *I* and *it* compared to other pronouns, and suggest lexical preferences for the omission-favouring initial position. Barring the conventionalised expression *thank you*, especially frequent with initial null subjects are different forms of the verbs GET, GO, and LOOK. Weir (2012) contrasts null subjects in spoken English with those found in diaries. While he analyses *diary drop* as instances of topic-drop, he concludes that the phenomenon is a purely phonological “left-edge deletion” in spoken language given its restriction to initial position. This assumption is, however, deduced from introspection rather than actual language data.

In lieu of authentic speech, the fictional representation of dialogue is used e.g. by Nariyama (2004), who, besides conversation, investigates TV scripts and casual letters. The most frequently omitted elements in all text types are *I* and *it*. Nariyama identifies four “triggers of subject omission”: anaphoric deletion (subject present in preceding sentence, can be reconstructed through linguistic information), deixis (subject can be reconstructed via non-linguistic context), dummy *it* deletion, and conventional expressions such as *gotta go* or *dunno* (Nariyama 2004: 252–253). In her analysis of spoken Northeast English, Bailey (2011) confirms the higher frequency of *I* and *it* as omitted forms, and reports a clear preference for overt *you*. Bailey corroborates her insights from sociolinguistic interviews with complementary questionnaires of acceptability judgments on various types of null subjects.

While the studies on English null subjects discussed above present recurring observations on types of null subjects encountered in different registers and varieties, none of these studies provides systematic quantitative evidence in the form of statistical evaluation of significance, or a multivariate factor analysis. Moreover, either only null, but not overt subjects are discussed, making it difficult to evaluate the validity of structural claims, or the quantitative basis concerning omission rates and types of null subjects is not made sufficiently clear. The following section discusses more representative, corpus-based studies with an explicitly variationist orientation.

2.3.2 Quantitative research on null subjects in English

Following years of almost exclusively generative approaches, researchers within usage-based and variationist paradigms have only recently paid attention to English null subjects, spawning a stunning number of studies in the last years. Factors identified as significant in these studies are used as a guideline and testing ground for the present investigation.

Of special interest here is the study by Teddiman (2011) who investigates null subjects in ICE-GB, including the sub-corpus “conversation (S1A)”, i.e. the same data as the present study. Furthermore, her study provides empirical evidence for hitherto rather anecdotal observations on the frequent co-occurrence of specific verbs and verb forms with null subjects in English. Measuring bidirectional association of utterance-initial verb forms with null subjects, she finds differing usage patterns in different genres of ICE-GB. ICE-GB conversations exhibit strong associations of the verb forms *sounds*, *looks*, *depends*, *feels*, and *seems* with null subjects. Additionally, she reports a preference of the negated auxiliaries *can’t*,

doesn't, *don't* and *didn't*, as well as non-negated *must* and *could* for subject omission (Teddiman 2011: 79).

A set of studies in the variationist quantitative paradigm is provided by Cote (1996), Harvie (1998), Leroux and Jarmasz (2006), Torres Cacoullos and Travis (2014), Travis and Lindstrom (2016), and Wagner (2012, 2018) on different varieties of North American English.

Cote (1996) uses the Switchboard corpus to extract approximately 400 subject tokens, with a 1:1 overt to zero token ratio. However, the method of extraction remains rather unclear. Harvie's (1998) analysis of Canadian English conversations uses a 2:1 overt to zero ratio, extracting the overt tokens from "contextually related adjacent clauses" (Harvie 1998: 18), a procedure also followed by Leroux and Jarmasz (2006) for their investigation of convergence effects between English and French in Canada. Cote (1996) focusses her investigation on effects on the discourse level. She finds a strong correlation of null subjects with utterance length, i.e. a pronoun omission favouring effect of short, especially one-utterance turns, and a higher amount of null subjects at turn boundaries. A similar effect for initial position is reported by Harvie (1998), who additionally confirms the universal effect of reference continuity for her data. While she describes a preference for null subjects to occur with negated forms, negation is not statistically significant; neither are the structural factors subject type (person, and referential vs dummy), clause type, and contrast in her analysis. Leroux and Jarmasz (2006) confirm the favouring effect of factors like coordination, initial position, and immediately preceding null token on English null subjects, but find no evidence for convergence between English and French with regard to null subjects, suggesting "to extensively study this variable across typologically distinct languages in the hope of defining the extent of its possibly universal nature" (Leroux and Jarmasz 2006: 12).

Most informative for the present investigation are the detailed variationist analyses provided by Torres Cacoullos and Travis (2014, 2015), and Travis and Lindstrom (2016) on US English, and Wagner (2012, 2018) on Newfoundland English. They are modelled on previous studies on Spanish and other canonical NSL, and systematically investigate the categorical and quantitative constraints on null subjects in the non-NSL English. Furthermore, they are distinguished by their transparent study design, making them a suitable starting point for the analysis in chapter 3.

Both Wagner (2012, 2018) and Torres Cacoullos and Travis (2014) restrict their analysis to first person contexts, arguing that

we cannot assume that all persons will pattern in the same way; first and second person, for example, have different information status from third person, and [first person singular] has a unique role in interaction as a reference to the speaker. (Torres Cacoullos and Travis 2014: 20)

However, given the frequent inclusion of wider pronominal contexts in the studies discussed in section 2.2.2, it is possible this particular selection is due to economic as much as conceptual reasons, an aspect Wagner (2012: 110–111) addresses openly. Rather than excluding contexts beside first person reference, the present study aims to operationalise these differences and integrate them into the analysis in chapter 3.

Torres Cacoullos and Travis transfer their insights from studies on Spanish subject realisation to the investigation of subject pronouns in spoken American English (see section 2.2.2). Torres Cacoullos and Travis (2014) analyse first person singular null subjects in 249,000 words from the Santa Barbara Corpus of Spoken American English (SBCSAE) with the aim to test the generative hypothesis that null subjects in non-canonical NSL are different from canonical null subjects beyond mere differences in frequency (Roberts and Holmberg 2010: 5). They find a total of 151 null subject tokens, compared to 9,000 overt first person singular pronouns, 6,600 of which occur “in the transcripts with at least one unexpressed *I*” (Torres Cacoullos and Travis 2014: 22). Apparently, almost a third of the texts, and accordingly a similar share of speakers represented in the data, show categorical overt subject realisation. This is strong evidence for the role of inter-speaker variation when investigating this non-standard grammatical feature for English. Torres Cacoullos and Travis exclude the invariant texts from their analysis, but do not address individual speaker variation any further. While they argue for the central role of accountability and point out the necessity for a comprehensive analysis to consider overt as well as null tokens, their database for the following analysis follows the approach to create a subset of data with an artificial 2:1 overt to zero pronoun ratio (including the preceding or following coreferential overt token). As argued by Wagner (2018: 3), it is not entirely clear how suitable such an artificial ratio is for the following statistical evaluation and the representativeness of their results. What is clear is the obvious economic advantage of this approach – representing the actual distribution of the variants in the data requires the coding of thousands of additional overt pronoun tokens.

Torres Cacoullos and Travis identify a highly restrictive variable context for English null subjects, i.e. the initial position in declarative main clauses, and coreferential coordination. For these contexts they consider the linguistic constraints coordination, priming, subject position in utterance, as well as speaker turn, subject continuity, polarity, semantic class of the verb, and tense (Torres Cacoullos and Travis 2014: 27). The latter five factors are not statistically signif-

icant, and, with the exception of subject continuity, not addressed in any more detail. While subject continuity shows a favouring influence on pronoun omission, it is not statistically significant, and not an independent effect in their analysis. Strong correlations of subject continuity are identified with the factors coordination and priming. Although these are not quantified as statistical interactional terms, *coreferential* coordination implies subject continuity by definition, as does priming, since it is defined as the “realisation of the previous coreferential [first person singular] subject” (Torres Cacoullos and Travis 2014: 25).

Besides the strong favouring effects of coordination, initial position and immediately preceding null token, Torres Cacoullos and Travis identify a set of “lexically specific” constructions for null subjects. Within the schema of coordinated verbs, the more specific expressions [I GO_{1SGI} and Ø VERB_{1SGI}] and [I VERB_{1SGI} and Ø QUOTATIVE VERB_{1SGI}] emerge as particular constructions, a finding similar to those on Spanish (e.g. Travis and Torres Cacoullos 2012, see also section 2.2.2). They conclude that

the crosslinguistic generalization lies in the existence of such constructions, in that particular constructions have also been revealed to play a role in subject expression in other languages, but at the same time, the constructions themselves are language specific. (Torres Cacoullos and Travis 2014: 32)

The investigation by Travis and Torres Cacoullos (2014) is concerned with stressed vs unstressed first person pronouns, a distinction commonly accounted for by similar principles of information management as the difference between overt and null realisation (see section 2.1.2). While they find that accessibility as an “information flow property” is crucial for stress on pronouns, it manifests differently than constraints on overt and null realisation: stress is not dependent on switch reference between adjacent clauses, but on referential distance (one or less vs two or more intervening clauses, Travis and Torres Cacoullos 2014: 384). The focus of Torres Cacoullos and Travis (2015) lies on the effects of English-Spanish language contact (see also section 4.6.2). Their results confirm the limited effect of reference continuity for English null subjects, especially compared to structural priming, and the language specific status of the initial position constraint (Torres Cacoullos and Travis 2015: 92). This contrastive approach to null subjects is developed further in a forthcoming paper on “variationist typology” (Torres Cacoullos and Travis *forthc.*).

Travis and Lindstrom (2016) provide the first detailed study on English to incorporate pronouns beyond first person contexts. They contrast different genres with regard to constraints on subject realisation in English, comparing third person singular human specific subjects in the SBSCAE with the English “Pear

Stories” (Chafe 1980). It is shown that subject omission rates are much higher in narrative than in conversational speech (22% null subjects in narratives vs 4% in conversation, Travis and Lindstrom 2016: 104). The findings of Torres Cacoullos and Travis (2014) on first person subjects, i.e. the restriction to initial position or coordination in declarative main clauses, are confirmed for third person. Despite the widely differing null subject rates in the two text types, they find that, as in Spanish, the constraint hierarchy remains constant across genres. It is thus only realisation rates that change between genres of the same variety, not the underlying grammatical system.

Wagner (2012, 2018) investigates the traditional L1 dialect Newfoundland English, known for its relatively liberal use of null subjects (Clarke 2004: 312). Analysing 34 interviews from the Pouch Cove corpus (approximately 280,000 words of conversation) yields 376 first person null subject tokens. First person is by far the most common context in this text type. Criticising the non-transparent token selection process of earlier studies, Wagner’s publications are the only studies so far to take into account all null and overt subject tokens in the data under investigation rather than an artificially created subsample. This results in more than 8,000 tokens analysed overall, with a subject omission rate of 4.4%. This rate is remarkable since Wagner does *not* include coreferential coordination, a main source of null subject tokens in most studies on English, and thus presents clear evidence for the higher frequency of null subjects in the Newfoundland dialect. This is also represented in a wider variable context than attested in the studies discussed above, as Wagner (2012) does report, and include in the statistical analysis, null subjects in subordinate clauses, and non-initial position. However, null subjects are rare in subordinated clauses, and not part of the follow-up analysis of the same data in Wagner (2018), raising the deletion rate to 7.2%.

In the following multivariate analysis, Wagner tests a wide variety of linguistic constraints known from previous investigations for her data, i.e. subject continuity (switch reference), clause type, turn position, turn length, preceding token, verb type and verb frequency. Additionally, the less well-known factor “verb phrase complexity”, found relevant mainly in studies on first language acquisition (especially Bloom 1990), is included. Analysing verb phrases as accumulation of sense units, she reports a clear correspondence of increasing complexity of the verb phrase with a higher likelihood for subject omission, accounted for by limited resources in language processing. Especially Wagner (2018) stresses the relevance of this factor, which might account for observations like the favouring effect of negation on null subjects, or be related to effects of verb types like auxiliaries or modal verbs.

Non-linguistic factors like speaker gender, age, and religion are not found significant for the Newfoundland data. Additionally, including speaker as a random intercept in a mixed-effects model only marginally improves model significance (Wagner 2018: 21). Besides verb phrase complexity, the linguistic contexts most decisive for subject realisation in Newfoundland English are preceding null token, initial position, and short turns. A slight favouring effect for overt first person pronouns is reported for perception verbs, such as *THINK*, *TELL*, *HEAR*, or *KNOW*, a finding in line with earlier research. No measurable effect is reported for verb frequency. While Wagner (2012) models persistence rather idiosyncratically by crossing the factor groups turn boundary, switch reference and preceding token, the latter factor takes clear precedence in influencing pronominal realisation, while the cross-linguistically influential factor switch reference is not significant in isolation; these tendencies are also reported by Torres Cacoullos and Travis (2015), and confirmed by Wagner (2018).

The preceding section has shown that there is growing interest in the quantitative investigation of non-canonical cases of null subjects, and increasingly so for the presumed non-NSL English specifically. Postulated categorical restrictions on null subjects in English, especially the supposed restriction to declarative main clauses, are possibly more appropriately modelled as strong probabilistic constraints. There are indeed very rare cases of embedded null subjects, which are additionally influenced by variety-specific preferences and aversions. Given the dominance of North American English in the presently available analyses, and the observed differences between US and GB with regard to null subjects, the detailed analysis of British speech data presented in chapter 3 offers the possibility to evaluate Biber et al.'s (1999), and Hageman and Ihsanen's (2001) observations on differences in usage and acceptability between US and British English. Beyond being an end in itself, given the role of British English as the historical superstrate of the Asian Englishes investigated, this clarification is also a necessary prerequisite for the comparative analysis in chapter 5.

Contrastive analysis of different languages and contact varieties stresses the importance of detailed structural analysis, as demanded by Torres Cacoullos and Travis:

The systematic quantitative analysis of variation in speech thus enables shared and language-specific patterns to be discerned. More such studies are called for to allow for the characterization of viable cross-linguistic generalizations on subject realization. (Torres Cacoullos and Travis 2014: 32)

Concerning structural constraints on English null subjects, cross-linguistic tendencies like referential continuity, and especially structural priming, are largely

confirmed to be valid for English as well, while results on verb semantics and polarity are inconclusive and possibly related to factors like collocational patterns, verb frequency, and verb phrase complexity. Due to the current lack of systematic investigation, it is difficult to evaluate the status of the structural factors subject person, and reference or specificity. Besides, the occurrence of lexically specific or conventional expressions favouring null subjects calls for further quantitative evidence.

Frequently made observations more specific to English include a strong preference of null subjects for coordination and utterance-initial position, and a predominance of omitted *I* and *it* compared to other pronoun forms. The present study aims to contribute to the clarification of language specific and cross-linguistically valid patterns of subject pronoun realisation, by investigating the manifestation of such structural constraints in a sample of British English conversation in chapter 3, followed by an investigation of the status of these constraints in contact varieties of English in Asia in chapter 5.

3 Empirical baseline: Null subjects in Spoken British English

Now you see most things in English are possible <,>

Some things are simply not as likely as others <ICE-GB:024#148–149:A¹>

Since to date there is no multivariate study on Standard British English available, the first step of this research is to establish the frequency and conditions of null subjects in spoken British English as a baseline. The aim is to provide transparent, reproducible results, to serve as the basis for further comparative analyses, e.g. with other L1 varieties of English, or different genres, which is missing from the research literature so far. In the following chapters 4 and 5, these findings will be matched with the three Asian contact varieties of English.

This chapter follows the tradition of variationist analysis in performing a quantitative statistical analysis of possible structural determinants of “near-synonymous” expressions, in this case the syntactic function of subject fulfilled by either pronoun or zero. The envelope of variation of the variable is circumscribed in section 3.1, the structural factors analysed are introduced in section 3.2., followed by a descriptive account of the data and an assessment of extra-linguistic factors in section 3.3.

The analysis tests predictions and linguistic factors proposed and established by earlier studies, as discussed in chapter 2, and their impact on the realisation of subject pronouns in British English. The statistical tool is logistic regression, conducted by the open source software R in combination with various dedicated software packages (for a complete list, see references). Its results are presented in section 3.4 for the full data set, and in section 3.5 for first person contexts specifically. Discussion and interpretation of the results is provided in section 3.6.

3.1 The envelope of variation: Coding decisions and examples

Defining the envelope of variation is the crucial first step underlying each variationist analysis (Tagliamonte 2006: 86–87). This section argues and illustrates what constitutes a null subject in this study, the variable context that is taken into account, and the operationalisation of null to overt subject ratios. The procedure consists of two steps: first, a priori exclusions are undertaken based on previous

¹ Sidney Greenbaum

empirical accounts of null subjects, followed by a detailed inspection of the data concerning the distribution of null subjects. The fact that the investigation in this chapter serves as the basis for a comparative analysis adds another level of complexity: several analytical categories are inspired by research on Asian languages and varieties; they are required for the comparison in chapters 5 and 6 (see also sections 2.2, 4.1, and 4.4). Therefore, the envelope of variation is circumscribed less orthodox than for an analysis of British English alone.

In accordance with the consensus in previous research (e.g. Li et al. 2012, Torres Cacoullós and Travis 2014), null subjects in this study are defined as the absent pronoun of a finite verb phrase. This excludes for example participles in non-finite sub-clauses, which do not require overt subjects in English (generative “big PRO”, Chomsky 1981: 56), but also instances of subject + verb or operator omission. Following the assumption by e.g. Torres Cacoullós and Travis (2014: 21), and Nariyama (2004: 243, but cf. Wagner 2012: 114–115) that subject + VP ellipsis behaves differently from subject-only omission, this excludes null subjects in subject + aux omission (3.1). Also excluded are overt pronouns without VP, a phenomenon encountered mainly in connection with copula deletion (3.2)

(3.1) A Are you running late nights working hard for exams or?

B Yeah

A Ø Burning the candle at both ends quite a bit, are you?

<ICE-GB:087#160-162>

(3.2) I Ø damn stupid lah. <ICE-SG:085#74:B>

The unit of analysis is the complete clause. Thus, subject pronouns in false starts and fragmentary utterances are excluded (marked **bold** in 3.3). Pronouns in repetitions are counted only once (the first instance in 3.4, see also Tagliamonte 2006: 93).

(3.3) I can't sleep properly ever, **cos I** when I go to bed at night

I I have to st my arm won't straighten you see. <ICE-GB:052#73:B>

(3.4) Well **it's** it's it's just so base. <ICE-GB:022#198:1:D>

Further contexts were excluded from the analysis because they are invariant in the (non-)occurrence of subject pronouns, and thus constitute categorical rather than probabilistic constraints on null subjects (Travis and Lindström 2016: 104), e.g. imperatives as canonical contexts for null subjects in English, and tag questions as invariant contexts for overt pronouns. Discourse markers like *I think* and *you know* are found invariant in ICE-GB after inspection of the data (see also Tagliamonte 2006: 90 on “formulaic utterances”). They are distinguished from

matrix clauses (3.5) via their position as parentheticals or postposed tags (3.6; see also Wagner 2012: 125).

- (3.5) Hopefully it will improve Neil a bit
but **I think** he's happier where he is now. <ICE-GB:025#285:B>
- (3.6) That was my big one of my biggest achievements **I think**
to do that sort of thing. <ICE-GB:028#147:A>

Unlike in the study by Torres Cacoullos and Travis (2014: 22–23), the comparatively marginal cases of null subjects in subordinate clauses and questions are part of the present analysis (see also sections 2.1.3 and 3.3). For one, they are not categorically invariant (see also Wagner 2012: 119), and moreover, as discussed in chapter 5, these contexts are part of the comparison with the Asian varieties. Wagner (2012) also includes subordinate clauses, but like Torres Cacoullos and Travis (2014) identifies questions as invariant syntactic context. This is probably due to the focus of both studies on first person pronouns – null subjects in questions are most commonly second person pronouns (Wagner 2012: 115).

While earlier studies on English are restricted to either first person pronouns, or do not include the full range of overt pronouns (or both, see also section 2.3.2), this study seeks to present the full picture of null subjects in English by including all persons. Overt pronouns were extracted (semi-)automatically, zero pronouns manually by reading through the transcripts and marking empty subject slots. Coding for the structural factors introduced in the following section was done manually by the present author, ensuring consistency in coding and interpretation, and covers the whole range of overt subject pronouns.

Rates of subject omission are calculated as the ratio of overt pronouns within the envelope of variation and zero, excluding invariant contexts and full NPs in subject position. Accountability is a crucial methodological principle of variationist analysis (see also section 2.2.1). However, most investigations of null subjects in English do not take the complete range of overt pronominal subjects into account, but rather operate on more or less clearly defined subsets, commonly a balance of 1:1, or 1:2 zero to overt pronouns, which is obviously a huge distortion of linguistic reality (see also section 2.3.2). While this approach is understandable from an economic point of view, it is not clear how reliable conclusions based on these skewed null subject rates are; the present study thus follows Wagner's stance in including the full set of pronominal subjects into the quantitative analysis (Wagner 2012: 131–132).

3.2 Analytic categories: Linguistic factors

A widely shared result of studies on null subjects is the crucial role of linguistic factors for subject realisation. Regional dialect and genre or style being constant, as is the case in the present study, structural constraints take clear precedence over social determinants in canonical, “radical” and non-null subject languages (see e.g. Travis 2007 for Spanish, Li et al. 2012 for Chinese, Wagner 2012 for Newfoundland English). While the exact circumscription of linguistic factors differs, a number of recurring, *de facto* standard categories have been established in numerous earlier investigations on null subjects in various languages (e.g. the summary in Bayley 2013 for Spanish, see also section 2.2.2). Most of these factors are also part of the main empirical models for this study, i.e. Wagner (2012), Torres Cacoullos and Travis (2014), and Li et al. (2012), with additions to test the influence of cross-linguistic tendencies and substrate-induced properties in the Asian varieties.

This section introduces the linguistic contexts that are analysed, clarifying their status in the present study and sketching the expected outcome. For the sake of transparency and replicability of results, the categories used in the present study are discussed in detail, and illustrated by examples from ICE-GB. The most common category, or factor level, is chosen as the reference level for a factor group in the statistical modelling by default, unless contrasts relevant for the purpose of the present study are represented better by a different selection (Helasvuo and Kyröläinen 2016: 278).

3.2.1 Coreferential coordination

One canonical context for null subjects in Standard English is the second subject slot in coreferential coordination (3.7; see also Biber et al. 1999: 156; on an alternative view of conjoined verbs as single clauses with two predicates, see e.g. Huddleston et al. 2002: 238). There is no unanimous agreement on the inclusion of this context, with Wagner (2012: 79) claiming that “[these examples...] are omnipresent and usually not classified as null subjects”. In contrast, Travis and Lindstrom (2016: 108) argue that the variable occurrence of pronouns in coordination remains unaccounted otherwise, concluding that “the variability in English subject expression that exists in both coordinated and non-coordinated contexts must be accounted for in order to obtain a full understanding of this phenomenon”. Null pronouns are indeed very common in these contexts; 59 out of 113 coreferential coordinations in the present data set appear without overt subject pronoun (3.7). Within the class of coordinated subjects, Torres Cacoullos and

Travis (2014: 28) identify temporal contiguity of the verbal action as a semantic factor favouring zero, whereas sequential situations favour overt pronouns (3.8).

- (3.7) If they become hungry they eat the book
and Ø read the uh <,> and Ø read the instructions on the baked tins.
 <ICE-GB:025#315:A>
- (3.8) Quite incredible we caught the evening boat
and we got there
and we changed some money. <ICE-GB:021#70-71:C>

Hendiadys, “[the realisation of] a single conceptual idea by two distinct constituents” (Hopper 2002: 146), is instantiated in English on a scale of more or less conventionalised expressions, the most lexicalised being Quirk et al.’s “pseudo-coordinations” *try and...*, *go and...*, etc. (1985: 979, see also Barth-Weingarten and Couper Kuhlen 2011: 279). Related to this are the “lexically specific” coordinations detected by Torres Cacoullos and Travis (2014: 31), which are responsible for a large share of null subjects in their data. They identify two kinds of verbal schemas. One kind evokes Hopper’s (2002: 152) “core hendiadic construction”, *go and ...*, with motion verbs as first constituent, [*I* GO_{1SGI} *and* Ø VERB_{1SGI}]. It is found both as fixed expression with *go* (3.9), but also with other inflected forms of GO (3.10). The other schema contains quotative verbs as the second conjoined element ([*I* VERB_{1SGI} *and* Ø QUOTATIVE VERB_{1SGI}]; 3.11).

- (3.9) And they go and Ø have a teabreak. <ICE-GB:027#157:C>
- (3.10) Uhm <,> I went and Ø watched the rugby. <ICE-GB:025#231:A>
- (3.11) And he rang and Ø said that she was very ill. <ICE-GB:028#126:A>

Torres Cacoullos and Travis (2014: 32) argue that the existence of lexically specific constructions favouring null subjects is valid cross-linguistically, but that the expressions themselves are language specific. One aim of the present study is thus to find out whether in a language internal comparison these co-occurrence patterns turn out to be variety specific, both with regard to a comparison of Standard Englishes (ICE-GB vs US English, analysed in Torres Cacoullos and Travis 2014) and contact varieties, as addressed in chapter 5.

Coreferential coordination is one of the most significant factors in Torres Cacoullos and Travis (2014), and is also expected to predict most occurrences of null subjects in British English. All clauses were thus coded for presence or absence of coreferential coordination, with absence, i.e. no coordination, as the reference level.

3.2.2 Clause type

One of the strongest structural claims by generativists about English null subjects is that they are a root phenomenon, i.e. they occur in declarative main clauses exclusively (e.g. Haegeman and Ihsane 1999: 126, see also section 2.3.1). However, this supposedly categorical constraint has repeatedly been refuted in the analysis of natural language data (e.g. in specific registers like diaries, Haegeman and Ihsane 2001, and in regional dialects, e.g. Ihalainen 1991, Wagner 2012). Especially second person zero can be encountered in questions, most likely in connection with operator drop, in examples like *Know what I mean?* (Biber et al. 1999: 1105; see also Wagner 2012: 115).

All clauses are marked as either declarative main clause (m), subordinate clause (s), or question (q). In an attempt to gain more insight into the nature of null subjects rather unexpected in this syntactic context, subordinate clauses are further distinguished into the categories general (3.13), relative clause (3.14) and *if*-clause (3.15). However, the low overall token number of null subjects outside declarative main clauses (3.12) makes these finer categories untenable; for statistical analysis they are thus collapsed into the more general category subordinate clause. As commonly noted, spoken language uses non-canonical question structures, e.g. without operator or marked solely by intonation (see Biber et al. 1999: 211–212); since examples like *You know what I mean?* are a common and acceptable form for questions in informal speech, instances like (3.16) are also included as questions.

- (3.12) Ø Feels alright now. <ICE-GB:089#115:A>
- (3.13) Ø Don't think Ø makes that much difference. <ICE-GB:022#291:A>
- (3.14) They had two spinster ladies Ø lived there called Miss Mahon.
<ICE-GB:028#225:A >
- (3.15) Or if Ø wanted to be a prison officer for instance. <ICE-GB:084#190:B>
- (3.16) What do you think the term anamnesis means?
Ø Got any idea? <ICE-GB:053#133:A>

The overall occurrence of the clause types distinguished here varies significantly across different persons, for both overt and zero pronouns. As shown in Figure 3.1, first person pronouns in general are underrepresented in subordinate clauses, and especially questions. A Chi-square test reveals that the differences between persons in clause types are highly significant ($\chi^2 = 307.61$, $df = 4$, $p < .01$). That Torres Cacoullos and Travis (2014) confirm the categorical restriction of null subjects to declarative main clauses for spoken American English is thus possibly a mere concomitant of their focus on first person pronouns.

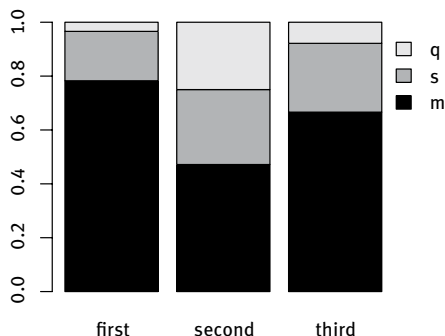


Figure 3.1: Percentage of clause types by person in ICE-GB conversations

As the most common context for both overt and null pronouns, main clause is the reference level for this factor group.

3.2.3 Position

Another common observation about English null subjects is their preference for utterance initial position (e.g. Biber et al. 1999: 1105–1106, Cote 1996, Harvie 1998, Thrasher 1974; see also section 2.3.2). In canonical sentences, English subjects occupy the preverbal position, quite commonly at the very beginning of the utterance, shifting the informative weight to the right. In spoken language, this is mirrored by reduced prosodic weight on the left edge (Quirk et al. 1985: 895–899). Due to the prosodic weakness of this initial position, subject omission in spoken language has been interpreted as an exclusively phonetic phenomenon (see e.g. Weir’s 2012 discussion of “left-edge deletion”).

Subject position here refers to the word number in an utterance. While this factor can theoretically be modelled as a continuous numerical measure, in reality Wagner’s (2012: 120) positions 1, 2, 3, and >3 are clearly sufficient. Non-initial subjects are usually preceded by interjections (3.18) or adverbials (3.19). A correlation between coordination and position >3, which is otherwise rather unlikely for an SVO language like English, is expected.

(3.17) Ø Can’t afford it. <ICE-GB:030#62:C>

(3.18) OK Ø got it. <ICE-GB:026#96:A>

(3.19) And then Ø went out to the theatre. <ICE-GB:025#235:A>

(3.20) He runs and Ø holds two poles. <ICE-GB:085#125:A>

Initial position (3.17) is the default for overt and zero subjects in English and thus the reference level. Two further structural contexts related to the factor group position are introduced below.

One factor that is at least partly related to position is turn boundary. Null subjects are discussed as a turn taking device, e.g. by Oh (2005, 2006). How this factor presents depends very much on the nature of a conversation, i.e. the degree of interactivity and the number of speaker changes. Turn boundary is a binary factor group.

Also on the level of conversational turns, short turns have been shown to favour null subjects e.g. in Cote (1996: 89) and Wagner (2012: 120). Turn length is measured in number of clauses within a speaker turn. Originally coded as a continuous factor in the present study, the factor levels were conflated to a four-way distinction (turns of 1, 2, 3, or 3+ clauses before a speaker change).

3.2.4 Person

As elaborated in section 2.1.1, different persons fulfil widely different functions in discourse. The most detailed studies on English null subjects so far have only addressed first person pronouns, presupposing that constraints are different for different persons a priori (Wagner 2012, Torres Cacoulios and Travis 2014; but cf. a recent study by Travis and Lindstrom 2016 on third person pronouns). To gain a more comprehensive understanding of null subjects in English, the present study includes all three persons.

Restricting the scope of the investigation potentially disguises issues like varying co-occurrence patterns of individual verbs or verb classes with different persons (Wagner 2012: 126–127). This leads to a skewed picture: factors found valid for individual persons are possibly mere artefacts of lexical variation, rather than truly syntactic variation. Supposedly general tendencies drawn from analyses of first person exclusively might thus not be valid beyond this context. Moreover, investigating the potential role of verbal marking for subject realisation is only possible with the inclusion of third person singular pronouns.

A further issue facilitated by the inclusion of all persons, which is presumably central for the varieties in chapters 4 and 5, is specificity and referentiality. While first person subjects are almost exclusively used as referential pronouns for animate subjects – usually discourse participants – second and third person pronouns show more variable usage. Second person pronouns can be used, like first person, for animate referents, but also with generic reference. Third person pronouns have the most diverse forms as well as functions. They can be referential, both internal and external to the discourse situation, or non-referential.

While number is a commonly used category in studies on null subjects, it is not included in the present investigation due to the inclusion of non-referential pronouns. Further issues with reference are addressed in the next section.

Several studies have found a favouring effect of first person for zero, and of second person for overt subjects (Bailey 2011, Flores-Ferrán 2007). The goal here is to test whether person itself is a relevant factor, but also to see to which degree constraints established for English first person subjects in Wagner (2012, 2018) and Torres Cacoullos and Travis (2014) are valid for second and third person pronouns as well.

3.2.5 Specific reference

Referential status and specificity of a pronoun are closely intertwined with the preceding factor. Specific referential subjects refer to concrete persons or entities, they have an identifiable individual or set of referents. For first and second person pronouns, referential is identical to animate, or human reference. There is no instance of generic *we* in the present data set. For second person pronouns, the distinction is made between referential (3.21) and generic usage (marked bold in 3.22); third person is distinguished between referential (3.23) and non-referential, more specifically existential (marked bold in 3.24), dummy pronoun (3.25), and generic usage (3.26). Weather predicates, another common source of non-thematic subjects, are surprisingly rare in the British data set, and none of them occurs with a null subject.

- (3.21) Have you got an example in mind of a text grammar? <ICE-GB:024#99:A>
- (3.22) There is a theatre in war you know [...]
You have the principal players the scenery you know the whole thing.
 <ICE-GB:052#14-17:B>
- (3.23) I've almost finished [this book]_i
 [It]_i 's very interesting. <ICE-GB:053#98-99:A>
- (3.24) Well books **there** were books but there were no magazines
 no newspapers no
Ø Weren't as many. <ICE-GB:084#6:B>
- (3.25) **It** would be very useful for the project if he did that. <ICE-GB:024#94:B>

- (3.26) 1 But there's a lot of people you get who who won't accept that
 aren't willing to argue
 2 **They** knock on your door on Saturday mornings
 3 and Ø get you out of bed
 4 and Ø say
 5 do you want to buy our pamphlet yeah. <ICE-GB084#117–119:A>

For the statistical analysis in section 3.4, the factor specific reference is recoded as a binary distinction between referential (as the reference level) and non-referential. Generic and non-referential pronouns are less specific semantically than referential subjects (or even semantically empty, as in the case of existentials; see also section 2.1). A decrease in semantic content is potentially associated with less overt linguistic material cross-linguistically. It is thus expected that non-referential pronouns favour zero expression, although more so in the Asian varieties than in Standard English (see also sections 4.1 and 4.2).

3.2.6 Switch reference

Switch reference, i.e. the change of subject referent from the immediately preceding subject, is one of the most widely acknowledged factors in studies on null subjects (see e.g. Bayley and Pease-Alvarez 1997, Otheguy et al. 2007, Otheguy and Zentella 2012, Silva-Corvalán 1994 and Torres Cacoullos and Travis 2010 on Spanish, Harvie 1998 and Wagner 2012 on English). Due to its basis in general principles of cognitive processing, its role in subject realisation is supposedly universal.

Addressing issues raised by Wagner (2012: 73–74), the initial coding of this factor indicated the amount of both functional and formal identity or switch from the arguments of the preceding clause, regardless of speaker changes. Similar classifications are used by e.g. Meyerhoff (2000, on Bislama), Jia and Bayley (2002), and Li et al. (2012, both on Chinese). Concerning functional or referential switch, a tripartite distinction is introduced:

- reference maintenance: the subject referent is identical with the subject of the preceding clause;
- partial switch: the referent is switched from subject of the preceding clause, but present in a different argument, e.g. the direct object; or shares partial reference with the preceding subject, e.g. the switch from *we* to *I*;
- full switch.

Contrasts marked in the form of the preceding item are:

- preceding noun;
- preceding pronoun with different form (e.g. a switch from *you* to *I* after a change of speaker, but potentially with the same underlying referent);
- preceding pronoun with identical form;
- preceding zero.

Compared to new referents, which tend to be expressed with overt subjects, reference maintenance or continuity is deemed more accessible cognitively. It therefore requires less linguistic coding, consequently favouring omission (as e.g. predicted by accessibility theory, see section 2.1.2). The strongest triggering effect for overt pronouns is reported by Li et al. (2012) for partial rather than full switch. While they do not further discuss this finding, this is possibly due to an increased potential for ambiguous reference in these contexts.

The classification containing both functional and formal considerations turned out too fine-grained for the data analysed here; especially partial switch is quite rare and contains too few tokens to allow for further distinctions. The factor switch reference was thus simplified to the three levels reference maintenance (m), partial switch (p), and full switch (s, as the reference level). The influence of preceding forms is introduced as the separate factor persistence in the following section.

3.2.7 Persistence

Lexical and structural persistence have been shown to play a role in syntactic variation in numerous experimental and corpus-based studies (e.g. Gries 2005 on dative alternation, Szmrecsanyi 2005, 2006 on future markers, comparative markers and genitive constructions). Apparently, overt as well as null forms trigger further occurrences, and even clusters of the same variant in neighbouring utterances (see e.g. Poplack 1980 on redundant plural marking in Puerto Rican Spanish, Poplack and Tagliamonte 1993 on verb phrase marking in earlier Black English). Persistence is also observed to influence subject pronoun realisation, and can be considered a universal factor (e.g. Travis 2005, 2007 on Spanish, Torres Cacoullos and Travis 2014 on US English, see the discussion in Wagner 2012: 178–189 for a comprehensive account of persistence effects from different methodological perspectives). Unlike accessibility, persistence, or priming, is considered a quasi-mechanical effect rather independent of reference continuity by Torres Cacoullos and Travis (2014: 25) and Wagner (2012: 189–190). It therefore includes both referential and non-referential pronouns. Its measure is the form of

the immediately preceding subject token, either as full NP (3.27), pronoun (3.28), or zero.

- (3.27) This fish had swallowed the fly before it died. <ICE-GB:055#223:B>
 (3.28) I had sardines once with a fly on it
 I was just about to tuck into it
 and I noticed this great fly soaked in tomato sauce
 <ICE-GB:055#221–222:B>

While the factor persistence as modelled here only takes into account the immediately preceding token, there are several instances of multiple null subjects clustering in immediate proximity, providing further evidence for the contextual influence of zero tokens, even beyond subject continuity (3.29).

- (3.29) 1 Well I really want to get it finished by Christmas
 2 I've got so much history of art to do
 3 Ø_i Got the draft
 4 And uhm so I'm to be finished by Christmas
 5 Ø_i Came to January
 6 Oh Ø_i really want it finished by February definitely.
 <ICE-GB:084#218–224:C>

Preceding zero is thus expected to trigger further instances of null subjects. Since the huge majority of subjects are pronouns, preceding pronoun is the most common context and constitutes the reference level for this factor group.

3.2.8 Verb phrase

While an effect of the finite verb phrase on subject omission is a common insight of numerous studies, its conditions have been modelled differently, with several interdependent measures, including verb semantics (Helasvuo and Kyröläinen 2016, Otheguy et al. 2007, Silva-Corvalán 1994, Travis 2007, Travis and Torres Cacoullos 2012), verb phrase complexity (Wagner 2012, 2018), morphological marking (Tamaredo and Fanego 2016), and lexical frequency (Bayley et al. 2013, Erker and Guy 2012).

In canonical null subject languages morphological marking is highly influential. This is not only due to identifiability of the omitted subject by person and number marking, but also by a condition that has been termed “discourse connectedness”, referring to stable person and number, as well as tense and mood

marking of a verb phrase (Bayley 2013: 18). Due to the low amount of inflectional marking in English, this measure is not considered here.

Verb phrase complexity, based on cognitive processing complexity, is modelled as the number of “sense units” encoded synthetically or analytically in the verb phrase (Wagner 2012: 129–130). It is found significant for Newfoundland English by Wagner (2012, 2018). In the present study, verb type and Wagner’s VP complexity correlate, as the most complex types of VP obligatorily contain auxiliary and / or modal verbs in English.

The more prominent factor across previous studies, however, has been the semantic type of the verb associated with null subjects. Clear contrasts are found between the behaviour of lexical verbs (l) expressing actions (verbs of motion, etc., e.g. *come, go, make, take, ...*) and psychological verbs (p) (e.g. verbs of perception, mental state, expression etc., *think, know, guess, mean, see, hear, ...*), with the latter favouring overt subjects (Wagner 2012: 127). Two further verb classes investigated here are the semantically weak primary auxiliaries (x) *be, do, have*, and the modal auxiliaries (m).

Lexical verbs, as the most prototypical verb type semantically, are used as the reference level for this category. Auxiliary and modal verbs possibly favour null subjects, not least due to their role in syntactic and processing complexity of the verb phrase. Psychological verbs are expected to trigger overt subjects. The stronger tendency for overt subject pronouns with these verbs especially for first person contexts is explained by a subjectivity effect and the epistemic role of this verb class (Travis 2007: 116). Based on the cross-linguistic findings for this factor, these verb type effects can be considered universal.

In the present data set, 15 verbs meet the criteria defined for “high-frequency” lexemes by Erker and Guy (2012: 536) as accounting for 1% or more of the tokens in the corpus. Even for this comparatively small data set, the frequencies of individual lexemes approximate a Zipfian distribution (Figure 3.2). While the effect of high-frequency verbs as a sub-class found by Erker and Guy (2012) could not be confirmed so far (e.g. Bayley et al. 2013), the role of single high-frequency verb tokens, and their potential skewing effects within the broader verb categories here, cannot be ignored (see also Torres Cacoullos and Travis 2010: 194 on the effect of the high-frequency constructions (*yo sé* ‘I know’ and (*yo creo* ‘I think’ on the class of psychological verbs in Spanish). The role of individual verb tokens favouring null subjects is addressed in sections 3.3.3 and 3.4.3, and these collocates are expected to turn out language, or even variety specific in the comparison presented in chapter 6.

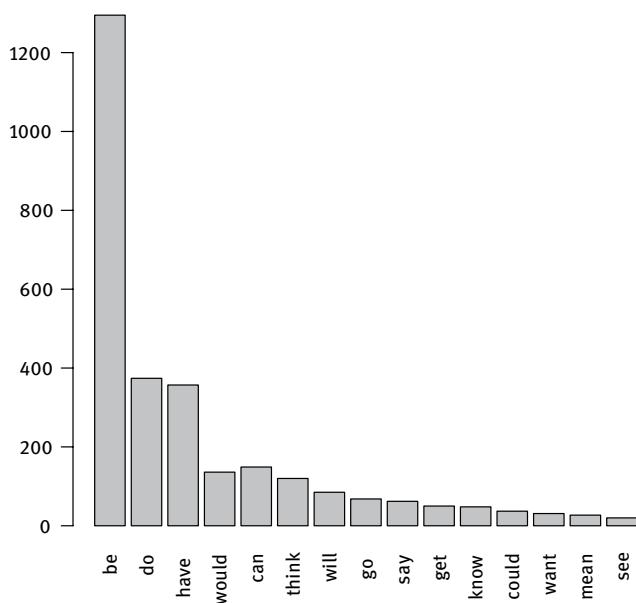


Figure 3.2: Raw frequencies of high-frequency lexemes in ICE-GB conversations

An overview of all linguistic factors investigated, including their reference levels and raw token frequencies for all contexts, is provided in section 3.3.4, Table 3.3.

3.3 Overview ICE Great Britain: Descriptive statistics and methodology

This section provides an overview of the data structure, a summary of raw frequencies and descriptive statistics, followed by an introduction of the statistical method used for analysis, binary logistic regression.

3.3.1 Distribution of null subjects in the data set

Twenty files from the “informal conversation” register of ICE-GB, i.e. approx. 40,000 words, are subjected to detailed statistical analysis. Including all pronominal subjects in variable contexts results in 4,272 tokens, with a rate of 3.04% zero pronouns (4,148 pronouns, 130 zero). A broad count on the remaining 70 transcripts of direct conversations within the sub-corpus ICE-GB:S1A reveals this

distribution to be similar across the files within this register. The null subject rate here lies between that reported in Torres Cacoullos and Travis (2014: 22, estimated roughly 2% for first person singular in US English conversations), Travis and Lindstrom (2016: 103, roughly 4% for third person singular in US English conversations) and Wagner (2012: 100, 4.4% for first person singular and plural in Newfoundland English interviews), although the latter study was conducted on a variety known for rather liberal use of this feature.

An alternative measure is null subjects per 1,000 words (approximately 3.25 in the present study, 4.56 in Wagner 2012 for all persons in Newfoundland English). Teddiman's (2011) investigation is also based on the ICE-GB conversation files, but only counts utterance initial zero. She reports 1.44 initial null subjects per 1,000 words (Teddiman 2011: 77), a number fairly close to the 1.2 initial null subjects per 1,000 words found here. The overall null subject rates found for the present data set are thus within the spectrum of preceding studies on the phenomenon in spoken English.

Each of the 20 files contains between one and up to twelve zero tokens, although most files are in the range of three to eight zero tokens (per approximately 2,000 words; Figure 3.3).

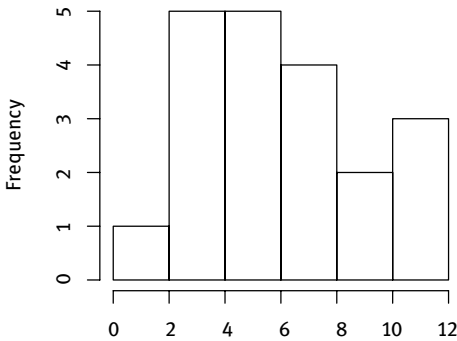
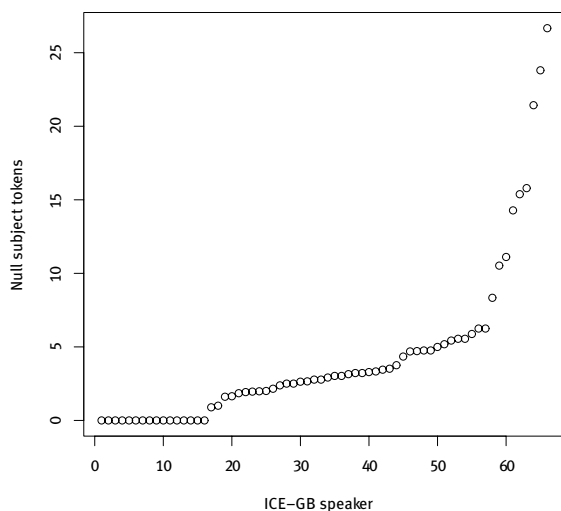


Figure 3.3: Histogram of null subjects per file in ICE-GB conversations

More remarkable is the fact that there are considerable differences between speakers. The role of individual speakers in modelling language variation has gained importance in recent variationist investigation (Wagner 2012: 133–134). Figure 3.4 shows the individual drop rates as percentage of null subjects as defined in section 3.1, each dot representing one speaker (ordered by null subject rate).



31 speakers are male, 19 female. Figure 3.5 shows a comparison of null subject rates per speaker between these two groups.

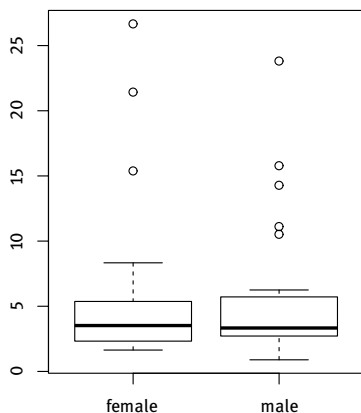


Figure 3.5: Percentage null subjects by gender ICE-GB

Boxplots are a convenient way to present such comparisons, as they provide several descriptive measures in graphical form. The bold line represents the median value of null subject rates in either group, the boxes represent the range within which the middle 50% of the observations fall (the interquartile range [IQR]). The whiskers represent values within the limits $1.5 \times \text{IQR}$ respectively. Outliers, i.e. speakers exhibiting values beyond this limit are represented by the single dots (Field et al. 2012: 145). Given that the median value for the amount of null subjects in each group is practically identical, and the range and number of outliers are similar for both groups, it is clear that speaker gender is unlikely to have an influence on subject realisation; this is also confirmed by a Chi-square test ($\chi^2 = 0.59$, $df = 1$, $p = .44$).

The ICE-GB metadata provides four age groups. Speakers are represented rather evenly across groups, with the exception of the oldest group 66+. It contains only one speaker with variable subject usage, who is thus subsumed under 46+, resulting in three age groups (18–25: 14 speakers, 26–45: 18 speakers, 46+: 14 speakers). A great amount of variance concerning null subject rates is encountered in the age group 18–25, the other groups show more homogeneous behaviour (Figure 3.6).

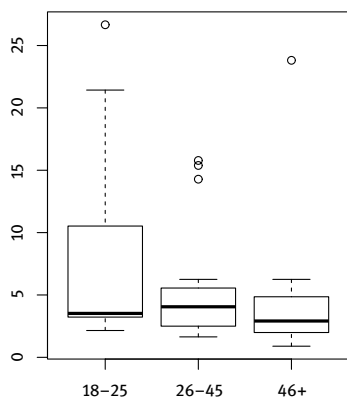


Figure 3.6: Percentage null subjects by age group ICE-GB

Still, the three groups are very close in their median values, and a Chi-square test comparing the distribution of null subjects is not significant ($\chi^2 = 1.26$, $df = 2$, $p = .53$). These results confirm earlier findings on the lesser significance of extra-linguistic factors for subject realisation.

3.3.3 Collocation

Individual speakers have been shown to influence variation; similar observations are made for individual verbs (Tagliamonte 2006: 223). While it is conceptually not entirely clear how empty categories can be explained in terms of co-occurrence patterns compared to overt elements, empirically the preference of individual lexemes for null subjects is evident (e.g. Teddiman 2011, Travis 2007).

Different measures are used to weigh the association strength of co-occurrence patterns (for an overview, see Levshina 2015: chapter 10). The affinity of different verbs to zero subjects can be measured by “attraction” and “reliance” scores (Schmid 2000). *Attraction* is the relative frequency of a verb in a given construction compared to the overall uses of the construction in the corpus, while the *reliance* score bases the comparison on all uses of the respective verb (Levshina 2015: 228). A rather infrequent verb with a high relative amount of null subjects will thus exhibit a high reliance, but a low attraction score, while a high-frequent verb with an accordingly high absolute number of null subjects, which is low proportionally to its overall use, will show the opposite. Cases in point here are *PULL* (four occurrences total, two of them with zero = high reliance score) and *DO* (354 occurrences total, 20 of them with zero, of 130 overall zero in the corpus =

for repulsion (these values represent $-\log_{10}(0.05)$ and $\log_{10}(0.05)$, respectively). Table 3.1 shows verb forms significantly attracted to null subjects.

Table 3.1: Verb forms attracting null subjects ICE-GB

GB Verb Form	overt	zero	total	logpvF
<i>sounds</i>	10	5	15	4.246
<i>see</i>	8	4	12	3.475
<i>doesn't</i>	9	4	13	3.325
<i>pulls</i>	1	2	3	2.568
<i>go</i>	16	4	20	2.566
<i>went</i>	31	5	36	2.370
<i>gets</i>	2	2	4	2.276
<i>don't</i>	126	10	136	2.102
<i>get</i>	17	3	20	1.668
<i>can't</i>	48	5	53	1.666
<i>saw</i>	8	2	10	1.453
<i>got</i>	23	3	26	1.367
<i>goes</i>	10	2	12	1.304

Verb forms with the highest collocational strength confirm the results for verb lemmas presented in Figure 3.7, with some additional insights. Apparently, contracted negation, especially with the auxiliaries *DO* and *CAN*, has a favouring effect on null subjects. For modal and primary auxiliaries, it is only negated forms that show a statistically significant attraction. Due to the correlation of negation with these verb types, polarity is not investigated as an independent factor in the following analysis. The lexemes *GO*, *GET* and *SEE* are prevalent with null subjects in various forms (base, third person, past), while the highly attracted form *sounds* is the only significantly attracted form of the lexeme *SOUND*. This points towards different distribution of these two types: *GO*, *GET* and *SEE* are apparently used more flexibly in constructions favouring null subjects (e.g. coordination, see also section 3.4.3), while the high attraction value for *sounds* is more indicative of a formulaic pattern with an initial omitted element (3.30).

- (3.30) D I suppose it's to be free tempo, isn't it?
 B Ø Sounds funny to me. <ICE-GB:026#326–327>

This also explains the strong significance of this verb form in Teddiman (2011: 79), who only considers utterance-initial contexts. On the other hand, her strong colllexemes *looks*, *depends*, *feels*, *seems* are dropped by the inclusion of all con-

texts, whereas negated auxiliaries, especially *can't* and *doesn't*, are more robust collocates of null subjects.

Table 3.2 presents the verb forms on the other end of the scale, which significantly repel null subjects.

Table 3.2: Verb forms repelling null subjects ICE-GB

GB Verb Form	overt	zero	total	logpvF
<i>is</i>	157	0	157	-1.828
<i>'re</i>	162	0	162	-2.02
<i>'s</i>	544	0	544	-7.604

While most verbs block null subjects in ICE-GB, the only forms showing statistically significant repulsion are forms of BE, especially contractions (see also Torres Cacoullos and Travis 2014: 22). The role of individual verbs for specific constructions like coordination, or in co-occurrence with different persons, is addressed in section 3.4.3.

3.3.4 Summary: Subject tokens and linguistic factors

Table 3.3 provides an overview of the linguistic factors used for the multivariate analysis in section 3.4, including the raw token numbers and percentages for each subcategory. The first level of each categorical variable is the reference level of the respective factor.

Obviously, not all categories of null subjects are equally common. The contexts provided as strong predictors, or even as categorical constraints for English subject realisation, are indeed underrepresented, especially the categories subordinate clause and question (within the factor group clause), and non-initial position. Together with coordination, these factors are expected to be English-specific. In contrast, factors based on cognitive principles of information processing and semantic aspects, such as switch reference, persistence, and verb phrase, are supposedly universal. Person and specific reference are factors possibly boosted by the substrates of the varieties analysed in chapters 4 and 5, or specific to contact languages.

Table 3.3: Linguistic factors and token numbers ICE-GB

		# zero	% zero	# overt	% overt	# total
Subject		130	3.75	3,338	96.25	3,468
and	no	71	2.12	3,284	97.88	3,355
	yes	59	52.21	54	47.79	113
Clause type	main	120	5.20	2,189	94.80	2,309
	subordinate	8	0.98	810	99.02	818
	question	2	0.59	339	99.41	341
Position	1	48	4.21	1,091	95.79	1,139
	2	5	0.72	688	99.28	693
	3	4	1.36	291	98.64	295
	>3	73	5.44	1,268	94.56	1,341
Turn boundary	no	108	4.28	2,413	95.72	2,521
	yes	22	2.32	925	97.68	947
Turn length	1	12	2.72	429	97.28	441
	2	18	3.89	445	96.11	463
	3	12	2.65	440	97.35	452
	>3	88	4.17	2,024	95.83	2,112
Person	first	49	4.16	1,129	95.84	1,178
	second	13	1.84	695	98.16	708
	third	68	4.30	1,514	95.70	1,582
Specific reference	referential	102	3.91	2,506	96.09	2,608
	non-referential	28	3.26	832	96.74	860
Switch reference	switch	29	1.54	1,855	98.46	1,884
	partial	8	2.53	308	97.47	316
	maintenance	93	7.33	1,175	92.67	1,268
Persistence	pronoun	105	3.47	2,917	96.53	3,022
	NP	14	3.71	363	96.29	377
	zero	11	15.94	58	84.06	69
Verb type	lexical	66	9.94	598	90.06	664
	psychological	23	5.15	424	94.85	447
	auxiliary	26	1.41	1,820	98.59	1,846
	modal	15	2.94	496	97.06	511

Given the provenance of the different factor groups, they are thus divided into three sources of factors for the feature pool of the Asian varieties (Table 3.4).

Table 3.4: Sources of linguistic factors

Origin	Factor groups	Predicted factor level favouring null subjects
English	Coordination Clause type Position	and-y main clause initial
Contact	Person Specific reference	possibly third person non-referential
Universal	Switch reference Persistence Verb type	reference continuity preceding zero lexical verbs

These distinct categories are especially relevant for the comparison with and between the contact varieties of English in chapters 5 and 6.

3.3.5 Statistical method

With the help of logistic regression, a widely used statistical tool in sociolinguistics, the influence of the explanatory factor groups (or predictors) and the individual factors (alternatively: factor levels) on the realisation of the binary response variable is measured. The output of the logistic regression model compares the chances of the variant zero to be chosen in one type of linguistic context with the chances for the variant *pronoun* in the same context (Levshina 2015: 254, for more details see also Baayen 2008: chapter 6 and Field et al. 2012: chapter 8). In the model built here, all predictors are categorical variables. Within the different factor groups, each individual factor level is measured against the reference level. This setting is called *dummy* or *treatment coding*, the default for logistic regression in R (Levshina 2015: 259).

The method of (multiple) logistic regression can be conducted with various software tools. Using the open-source software R and specialised additional packages is increasingly common in sociolinguistic studies. The procedure is closely related to the “classic” variable rule (“Varbrul”) analysis with the dedicated software “GoldVarb” (e.g. Sankoff 2008), but provides various advantages, e.g. greater variability with regard to predictor classes, and higher flexibility by integrating open source extensions and updates, as argued by e.g. Helasvuo and

Kyröläinen (2016), Levshina (2015), Tagliamonte and Baayen (2012), Wagner (2012), and especially Johnson (2009).

Conducting logistic regression analysis with R and its extensions also provides the possibility of building random intercepts and random slopes into the model (Baayen 2008: chapter 7); this is especially useful for variation between individual speakers or verb tokens, which could affect the dependent variable, but are not considered key factors of variation (Gries 2013: 333–334). However, due to the relatively low frequency of the phenomenon investigated here, models with added random intercepts for “speaker” and “verb token” could only be computed with great loss of reliability (see Table B.1 and Table B.2 in Appendix B), hence the discussion of these aspects in the preceding sections.

The following section presents the results of the logistic regression analysis of the British English data, based on the data and linguistic factors described above.

3.4 Null subjects in ICE Great Britain: Logistic regression model

This section establishes the comparative baseline for null subjects in English by fitting a logistic regression model to analyse the influence of the linguistic factors on null subjects in British English. After an evaluation of the model and a description of the model selection process, the results of the minimal adequate model are presented. All analyses were performed with R Version 3.3.0, a list of additional software packages used is provided in the references. This model will then be compared to the model fits for the Asian varieties in chapter 5.

3.4.1 Model selection

The initial maximal factor model contains ten categorical predictors with a total of 30 levels, tested on the binary dependent variable subject realisation. Overt pronoun is the reference level for the dependent variable, the model thus measures the comparative likelihood of zero pronoun in different contexts. Table 3.5 shows the Wald statistics for the maximal factor model, i.e. the statistical significance for all factor groups.

P-values <.01 indicate high statistical significance. This is the case for four factor groups (coordination, position, persistence, and verb type). The factors clause, person, and reference still fall safely into the range of $p < .05$, the conventional cut-off for statistical significance. Turn boundary is clearly not relevant

for the observed variation. There are two borderline categories – turn length and switch reference – hovering close to the .05 limit. The (relative) lack of statistical significance for switch reference is somewhat surprising – switch reference is a well described factor in earlier cross-linguistic studies. However, the role of switch reference is less clear for English null subjects (see sections 2.3.2, 3.2.6).

Table 3.5: Wald statistics of factor group significance ICE-GB

Factor group	Chi-Square	d.f.	p
and	102.00	1	<.0001
Clause	6.36	2	0.0415
Position	23.08	3	<.0001
Turn boundary	0.38	1	0.5397
Turn length	7.30	3	0.0630
Person	8.96	2	0.0114
Specific reference	5.90	1	0.0151
Switch reference	5.38	2	0.0678
Persistence	10.02	2	0.0067
Verb type	46.41	3	<.0001
Total	279.47	20	<.0001

To determine the best model fit, the maximal factor model is evaluated with the help of a bidirectional stepwise selection procedure (step up / down analysis, Levshina 2015: 267). This means that each factor group is added (step up) or erased (step down) one by one. Each reduced model is compared to the full model with regard to the diagnostic measures *deviance* and the *Akaike Information Criterion* (AIC, Field et al. 2012: 316–318). The matching results for forward and backward selection show that the factor groups turn length and switch reference both improve the model, but only marginally so (details of the model comparison are provided in Table B.3 and Table B.4 in Appendix B).

Model selection requires a balance of interpreting seemingly objective diagnostic values, and conceptual considerations. Due to the theoretical relevance of switch reference as an underlying cognitive principle (which is not assumed for turn length to the same degree), it is kept in the model, in spite of its tenuous p-value. This elimination process yields the minimal adequate model (see Table 3.6) – the remaining eight predictors (with 24 factor levels) significantly influence the outcome of subject realisation. After a presentation of indicators to evaluate model fit, in the discussion of the minimal adequate model the influence of the different factor levels is assessed in detail, as are possible, and actually observed, interactions between factor groups.

3.4.2 Model evaluation

Several indices for model significance, or goodness of fit, exist, and the minimal adequate model for British English satisfies the most common criteria (Levshina 2015: 258–259; a more detailed overview is presented in Table B.5 in Appendix B). The *Model Likelihood Ratio Test* (LRT) indicates the general significance of the model as compared to the intercept-only model, i.e. a model without any predictors. With a p-value $< .01$, this model is highly significant ($LR \chi^2 = 395.45$, $df = 16$). A commonly used measure for goodness-of-fit of the model is the *concordance index C*, with a desired outcome of > 0.8 (Gries 2013: 304), which the present model achieves easily ($C = 0.87$).

Another indicator for the validity of logistic regression models is *Nagelkerke's R*, also called *pseudo-R²*, measuring the predictive power of a model from 0 (no predictive power) to 1 (perfect prediction, Levshina 2015: 259). An R^2 index of 0.39 is interpreted as the percentage of variation accounted for by the model (see e.g. Field et al. 2012: 317–318, Wagner 2012: 137). However, this diagnostic measure stems from linear regression originally; to which degree it is valid to evaluate logistic regression models is not resolved conclusively (see e.g. Levshina 2015: 259).

A serious problem for models with many predictors is overfitting, i.e. generating significance and too optimistic discrimination indexes through the sheer number of possible influencing factors. Overfitted models perform well on the given sample, but poorly on any new data beyond that. Their insights cannot be generalised, accordingly their scientific value is very limited. The model is thus validated by bootstrapping, i.e. repeated resampling from the observed data set (200 runs in this case, of which 142 successfully completed, the full output can be found in Table B.6 in Appendix B). The model is then fitted to these random new data sets. From the average of these test-runs, the ability of the model to accommodate new data can be measured (Levshina 2015: 166–167).

Overfitting is signalled by *high optimism levels*, indicating the amount of overestimation to be subtracted from the *naïve estimates*; this results in the bias-corrected index values, which are necessarily lower than the original. The slope optimism for the present model is 0.04, below the recommended 0.05. (Levshina 2015: 167). With a value of 0.85 (compared to the original 0.87), the corrected C is still above 0.8, indicating excellent model fit. The same is valid for the corrected R^2 (0.37 vs the original 0.39). The model is thus expected to generalise well beyond the data investigated here.

Table 3.6 provides a summary of the minimal adequate model. Following common convention, factor levels are marked for significance with asterisks,

based on their p -values.² The model does not exhibit unusually high standard errors of a factor level, which may point to insufficient data, or signal an interdependence with another factor (*multicollinearity*; more on this below, see also Field et al. 2012: 322).

The default setting for categorical variables in R is treatment coding, i.e. comparing the levels spelled out in the output to the respective reference levels of the individual factor groups, usually the most common realisation (as discussed in section 3.3.5 above). The approximation of factor strength and direction of effect as compared to the reference level of each factor, are indicated by the coefficients. These measures of relative factor weight are given as *log odds ratios* centred around zero. This means positive values show a favouring effect of the respective factor level on null subjects, whereas negative values represent a disfavouring effect, i.e. the tendency for overt pronoun realisation in the respective context (e.g. Levshina 2015: 260–263; on alternative scales, see Gries 2013: 299–301).

Table 3.6: Logistic regression model ICE-GB

	Coefficient	SE	Z	p-Value	
Intercept	-2.6127	0.3081	-8.481	<.0001	***
and: y	3.6451	0.3641	10.011	<.0001	***
Clause: subordinate	-1.1142	0.443	-2.515	0.0119	*
Clause: question	-0.071	0.7648	-0.093	0.9260	
Position: 2	-2.5866	0.5411	-4.78	<.0001	***
Position: 3	-0.8274	0.5909	-1.4	0.1614	
Position: >3	-0.6519	0.3091	-2.109	0.0349	*
Person: second	-1.0376	0.4081	-2.542	0.011	*
Person: third	-0.0103	0.2834	-0.036	0.9711	
Reference: non-referential	0.8244	0.3273	2.519	0.0118	*
Switch: partial	0.2913	0.4697	0.620	0.5352	
Switch: maintenance	0.6402	0.2765	2.315	0.0206	*
Persistence: NP	0.594	0.3828	1.552	0.1207	
Persistence: zero	1.4388	0.5194	2.770	0.0056	**
Verb type: psychological	-0.6673	0.3154	-2.116	0.0344	*
Verb type: auxiliary	-1.8788	0.278	-6.759	<.0001	***
Verb type: modal	-0.9201	0.3439	-2.675	0.0075	**

Some possible interactions between factor levels are addressed in section 3.2. Besides unusually high standard errors, *Variance Inflation Factor* (VIF) scores

² *** < .0001, ** < .01, * < .05, . < .1

can identify multicollinearity between predictors (Levshina 2015: 272–273). While logistic regression is rather robust to correlation, in its most extreme form, multicollinearity can lead to complete separation of the data set, leading to unreliable estimates of the affected predictors in the model (Helasvuo and Kyröläinen 2016: 281).

Table 3.7: Estimated variance inflation factors ICE-GB

Factor level	vif
and: y	2.1
Clause: subordinate	1.38
Clause: question	1.12
Position: 2	1.22
Position: 3	1.16
Position: >3	2.03
Person: second	1.35
Person: third	1.75
Reference: non-referential	1.75
Switch: partial	1.27
Switch: maintenance	1.59
Persistence: NP	1.33
Persistence: zero	1.06
Verb type: psychological	1.28
Verb type: auxiliary	1.32
Verb type: modal	1.2

To rule out multicollinearity, VIF scores should be below 3 (Helasvuo and Kyröläinen 2016: 281); while some categories, e.g. coordination or position, show slightly increased interaction values, judging from Table 3.7, multicollinearity is not a problem for this model. To test their influence on the model fit, several interactions between factor groups were built into the model and compared to the original model. The only interaction that significantly improves the model is between person and verb type (see also section 3.2.4; the full model including the interaction term can be found in Table B.11 in Appendix B; the implications of this interaction are discussed in section 3.4.3.8 below.).

The following sections discuss the contribution of the individual factor groups in detail. To ease comparison with earlier sociolinguistic studies, especially those on null subjects in spoken English by Wagner (2012, 2018), Torres Cacoullos and Travis (2014), and Travis and Lindstrom (2016), the discussion of the individual factors contains VarbRul/Rbrul style factor weights (Tagliamonte 2006: 141–143). These are calculated as probabilities of null realisation, based on

the log odds for factor levels achieved via sum coding, i.e. comparing the likelihood of individual factor levels with the group average (“factor weights”; Johnson 2009: 361–362). This scale is centred around 0.5, with values >0.5 indicating a favouring effect of the factor level on null realisation, while values <0.5 show a pronoun preserving effect of the respective context (Gries 2013: 301). Following variationist convention, the factor levels are ordered by decreasing factor weight.

3.4.3 Results by factor groups

Now that the statistical significance of the minimal adequate model has been evaluated, it is time to inspect the relative influence of the different factor levels.

3.4.3.1 Coreferential coordination (“and”)

The first factor in the analysis, coreferential coordination, is also the one that has by far the greatest favouring influence on null subject realisation – as was to be expected, given the status of this context as the only canonical variable realisation in Standard English (Table 3.8).

Table 3.8: Results for factor group coordination ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
y	1.823	113	0.522	0.861	72
n	-1.823	3,355	0.021	0.139	

Although conjunction is the most likely context for null subjects, and in fact almost 50% of all null subjects in ICE-GB occur in this context, omission is not obligatory, especially in sequential or discrete situations (3.31, see also section 3.2.1).

- (3.31) I was in the navy five days after I was eighteen
And I was on board a ship six days after I was eighteen.
<ICE-GB:028#242:A>

Coreferential conjunction is highly significant as a factor group (see Table 3.5), and its presence as a weighting factor level towards subject omission exhibits the highest factor weight. Its influence is so strong that a monofactorial model containing coreferential conjunction as the only predictor was contrasted with

the full model here (provided in Table B.7 in Appendix B); however, the full model does prevail significantly with regard to its explanatory power.

Generally, coordination is more common with lexical verbs as the second element. Torres Cacoulos and Travis (2014: 31) identify two “lexically specific constructions” for American English favouring null pronouns, i.e. [*I* GO_{1SGi} *and* Ø VERB_{1SGi}] and [*I* VERB_{1SGi} *and* Ø QUOTATIVE VERB_{1SGi}]. Indeed, instances like (3.32) and (3.33) are common in ICE-GB.

(3.32) And they go and Ø have a teabreak. <ICE-GB:027#157:C>

(3.33) If they come to us and Ø say ‘we have to leave’. <ICE-GB:027#62:B>

Figure 3.8 and Figure 3.9 provide the attraction and reliance scores of verb lemmas with the first and second verb slot in coreferential coordination, respectively. The primary auxiliaries HAVE and BE are overrepresented due to their extremely high token frequency; to support clarity, they are not represented in the graphs below.

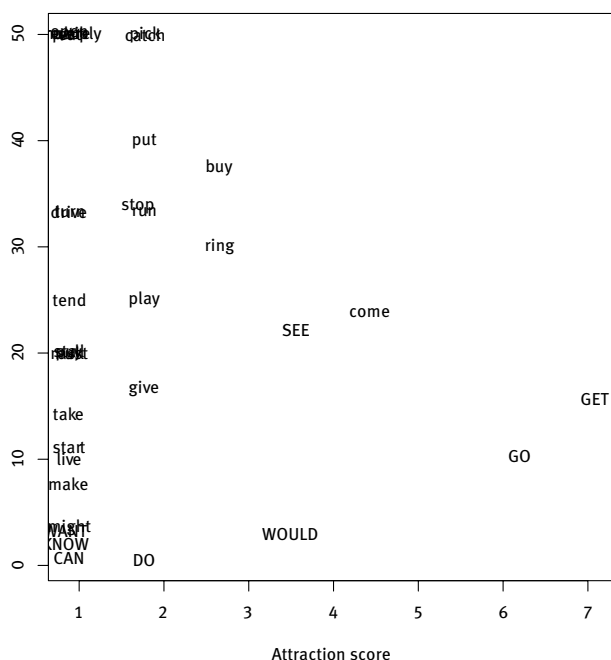


Figure 3.8: Attraction and reliance scores of verbs in position AND1 ICE-GB

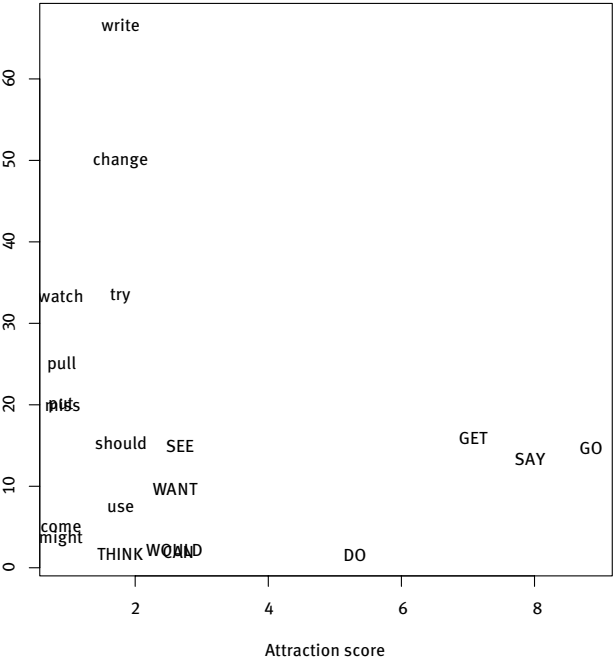


Figure 3.9: Attraction and reliance scores of verbs in position AND2 ICE-GB

The first verb slot in coordinations features a wider variety of verb tokens than the second. As predicted, the motion verbs *COME* and *GO* exhibit high reliance and attraction scores to this position within the coordinative construction. *SEE* and *GET* are other verbs likely to be found in this context.

While *GO* and *GET* are also strongly associated with the second verb slot in coordination, *SAY* as the most frequent quotative verb is only found in this position. The strong attraction of these verbs to the coordinative construction explains why they are not found as collexemes for null subjects in Teddiman (2011), who only considers utterance initial contexts (see also section 3.3.3). While collocational analysis is better suited for larger data sets than the present one, the tendencies observed in Torres Cacoullas and Travis (2014) concerning lexically specific constructions can be supported.

3.4.3.2 Clause

Although the assumed restriction to declarative main clauses is one of the strongest claims for English null subjects, clause is barely statistically significant as a

factor group (see Table 3.5). Compared to the reference level main clause, both questions and subordinate clauses disfavour null subjects, but only subordinate clauses show a statistically significant effect.

Table 3.9: Results for factor group clause ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
main	0.395	2,308	0.052	0.598	27
question	0.324	341	0.006	0.58	
subordinate	-0.324	819	0.010	0.328	

The disfavours effect of subordinate clauses is similar in weight to that found in Wagner (2012: 148, Table 3.9). Due to the low absolute numbers of null subjects in subordinate clauses, no specific types of subordinate clauses are further investigated. However, it should be noted that the tendency observed in Wagner (2012: 167) for “the most frequent combination [...] ‘mental stance verb’ + *that*-clause (complement or relative)” as a likely context for null subjects in subordinate clauses is not found in the present data set.

While null subjects outside declarative main clauses are not as impossible as the literature predicts, they are certainly rare enough to not consider them a systematic feature of spoken British English.

3.4.3.3 Position

Utterance initial position is not only favoured for subjects in general, but also for null realisation (Table 3.10). In comparison, all other utterance positions favour overt pronouns. Position is a highly significant factor group, with a strong effect of the reference level initial position.

Table 3.10: Results for factor group position ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
1	1.016	1,139	0.042	0.734	56
>3	0.365	1,341	0.054	0.59	
3	0.189	295	0.014	0.547	
2	-1.570	693	0.007	0.172	

The strong prediction for null subjects in initial position only might call for a mere binary distinction between initial vs non-initial position (which is the clas-

sification found in e.g. Torres Cacoullos and Travis 2014). This, however would bury a rather interesting effect that can be observed with regard to the more exact position of the omitted pronoun – it is least likely in second position, i.e. preceded by one lexical item only, whereas the disfavoured effect is less strong for position 3 and >3.

3.4.3.4 Person

Person as a factor group is only slightly significant in the model. The most common context overall is third person pronouns, which also exhibit the highest rate of subject deletion (Table 3.11).

Table 3.11: Results for factor group person ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
first	0.349	1,178	0.042	0.586	25
third	0.339	1,582	0.043	0.584	
second	-0.688	708	0.018	0.334	

The most significant result here is that second person is least common not only overall, but also for zero pronouns. Half of these are generic uses of second person pronouns, while they make up less than 20% of all second person pronouns (Figure 3.10).

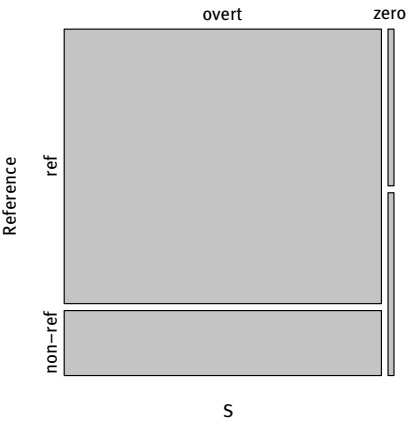


Figure 3.10: Referential vs non-referential second person pronouns ICE-GB

This result is in line with Bailey’s (2011: 29) acceptability rates for sentences containing null subjects, which are considerably higher for generic compared to specific second person pronouns.

One surprising insight is that first and third person pronouns behave more similarly, given their different functional load, compared to second person, which at first glance seems to share more functions with first person as designators of participants in the interaction. To enable a comparison with Wagner (2012) and Torres Cacoullos and Travis (2014), the behaviour of the subset of first person pronouns is discussed in more detail in section 3.5.

Wagner (2012: 126–127) raises the issue of different collocational patterns of the three persons. Given the preference of individual lexemes for null subjects, this aspect is possibly relevant for variation between persons. Apparently, first and second person have more distinct co-occurrence patterns than third person. THINK is associated with first person contexts very strongly, as are other psychological verbs such as KNOW, MEAN, REMEMBER, but also modal verbs CAN, WOULD, WILL and COULD (the interaction of person and verb type is further discussed in section 3.4.3.8).

3.4.3.5 Specific reference

A factor closely connected to person is that of subject reference. In the final analysis, the initially more fine-grained distinctions within non-referential pronouns are collapsed to a binary “referential – non-referential” distinction, with referential pronouns as the reference level.

Table 3.12: Results for factor group specific reference ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
non-referential	0.412	860	0.033	0.602	20
referential	-0.412	2,608	0.039	0.398	

There is a favouring effect of non-referential status for pronoun omission (Table 3.12), the factor group is, however, not very significant statistically (Table 3.5). Figure 3.11 shows interaction effects between the factor groups person and reference within the regression model. The visualisation shows all subject tokens distributed across the two factor groups person and specific reference, and their probability for subject omission (Levshina 2015: 269). The bold line shows the probability of the crossed factors, the light grey bars represent the confidence bands, while partial residuals are shown as dots above and below (Breheny and Burchett 2017: 3).

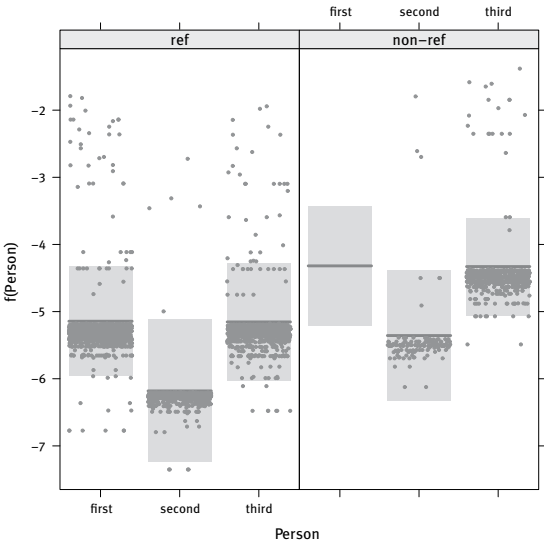


Figure 3.11: Interaction graph person by reference ICE-GB

As obvious from the graph, third person pronouns form the great majority of non-referential pronouns, while the combination of first person pronouns with non-specific reference (i.e. generic *we*) is not encountered in the present data set. A look at the subset of referential pronouns only shows a slight tendency towards singular null pronouns (4.3% vs. 2.9% for plural; however, this difference is found not statistically significant in a Chi-square test with $\chi^2 = 2.12$, $df = 1$, $p = .15$).

3.4.3.6 Switch reference

The role of switch reference for English null subjects is apparently limited (see also Wagner 2012: 147, Torres Cacoullos and Travis 2014: 27). The factor group is of restricted statistical significance, and barely made it into the final minimal model of the present analysis (Table 3.13).

Table 3.13: Results for factor group switch reference ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
maintenance (m)	0.330	1,268	0.073	0.582	16
partial (p)	-0.019	316	0.025	0.495	
switch (s)	-0.311	1,884	0.018	0.423	

The absence of a strong switch reference effect is rather unusual from a cross-linguistic perspective. Both partial and full switch have a slight favouring effect on overt pronoun realisation, but the effect is comparably weak, and accordingly the range of the factor groups is low. Switch reference obviously correlates with coreferential coordination, which automatically excludes the factor level full switch (Figure 3.12).

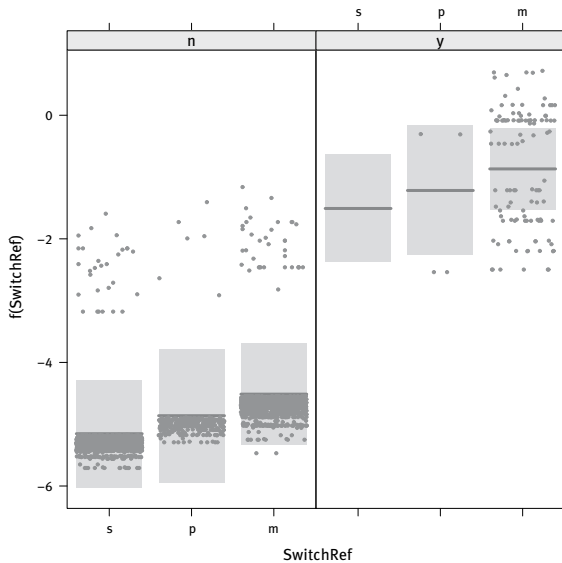


Figure 3.12: Interaction graph switch reference by coordination ICE-GB

Besides this interaction, the difference in the height of the horizontal lines between column n (no coordination) and y (coordination) in the graph illustrates the major effect of coordination on probability of subject omission.

3.4.3.7 Persistence

Due to the high frequency of pronouns in subject position, as compared to both zero and full NPs, preceding pronouns are the default case, and thus the reference level for the factor group persistence. Although it has been questioned conceptually whether zero elements can function as a structural primer at all (Wagner 2012: 182), persistence is a highly significant factor group (Table 3.14).

Table 3.14: Results for factor group persistence ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
zero	0.761	69	0.159	0.682	35
NP	-0.084	377	0.037	0.479	
pronoun	-0.678	3,022	0.035	0.337	

Similar to earlier studies, immediately preceding zero favours further subject omission, which has been described as a kind of mechanical priming effect (e.g. by Travis 2005, Travis and Torres Cacoullos 2012). On the other hand, full NPs, as more explicit clarification of reference, also have a favouring effect on subject omission compared to pronouns, although this effect is smaller, and less significant than for preceding zero. Figure 3.13 shows how null subject rates increase from preceding pronoun to NP to zero across all levels of switch reference.

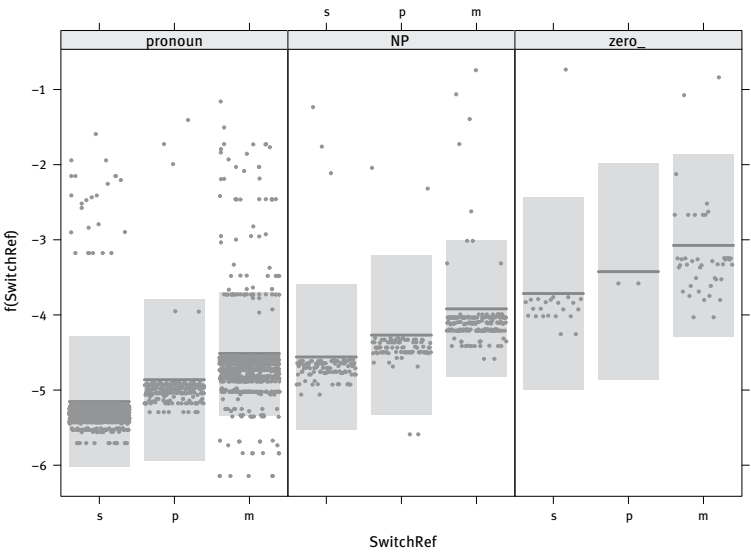


Figure 3.13: Interaction graph switch reference by persistence ICE-GB

It is still obvious that preceding zero shows a high correlation with reference maintenance. Partial switch is most common from preceding lexical NPs, which is due to the fact that these constitute a different argument of the preceding clause most commonly. Especially objects usually serve to introduce new information into the

discourse (see also sections 3.2.6 and 2.1.2). The combination preceding zero with partial switch is highly uncommon.

The strength of the favouring effect of a preceding zero subject, compared to the weaker effect of reference maintenance on subject omission, supports the interpretation of persistence as a mechanical rather than a semantic process.

3.4.3.8 Verb phrase

As discussed in section 3.2.8, the influence of the verb phrase on subject realisation can be modelled in different ways. Here, two (somewhat complementary) approaches will be addressed, based on semantic and syntactic features of the verb, and complexity of the verb phrase.

Wagner (2018) stresses the influence of *verb phrase complexity* for subject realisation. Her model also includes verb type, and it is not entirely clear how she handles the obvious correlation of higher complexity with the verb type auxiliary. For the present data, verb type offers more explanatory value, as can be seen from the comparison with a model containing verb phrase complexity as a factor group (see Table B.8 in Appendix B). Verb type is one of the most significant factor groups in the model discussed here (Table 3.15).

Table 3.15: Results for factor group verb type ICE-GB

factor	logodds	tokens	% zero	centred factor weight	range
lexical	0.867	664	0.099	0.704	44
psychological	0.199	447	0.051	0.55	
modal	-0.054	511	0.029	0.487	
auxiliary	-1.012	1,846	0.014	0.267	

Lexical verbs are clearly the most favourable context for null subjects, followed by psychological verbs. There is a large decline in null subject rates for both types of auxiliaries, especially primary auxiliaries BE, DO and HAVE (this is also found by Wagner 2012: 154). The disfavouring impact of primary auxiliaries on null subjects is one of the strongest effects in the whole model. Given their extraordinary frequency in discourse, the low null subject rate of these three lexical items significantly lowers the overall percentage of subject omission. This effect is also described by Bailey (2011: 39–40). One factor is the common tendency for cliticisation for BE and HAVE, which blocks subject deletion (Torres Cacoullos and Travis 2014: 22, see also Table 3.2 in section 3.3.3).

Figure 3.14 shows the expected correlation of different persons with different verb types, especially a higher co-occurrence of first person with psychological,

and third person with lexical verbs (see e.g. Wagner 2012: 126–127). It also shows a comparatively high co-occurrence of second person zero with psychological verbs, and the hindering effect of the combination third person with primary auxiliaries on null subjects.

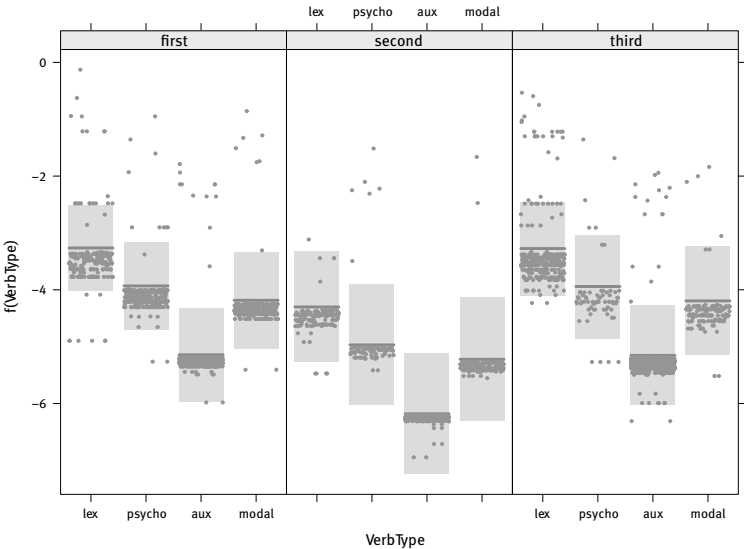


Figure 3.14: Interaction graph verb type by person ICE-GB

The interactional patterns between person and verb are quite obvious from Figure 3.14. However, it also reveals the non-attestation of second person zero in combination with modal auxiliaries. This gap makes the integration of this interaction into the model computationally difficult, in spite of its conceptual value and its attested improvement of explanatory power (the model including the interaction *person : verb type* is provided in Table B.11 in Appendix B).

To test the influence of individual lexemes within verb classes, a model with *verb token* as random intercept was computed (glmer). The effects of factor levels primary and modal auxiliary are decreased in this model (see Table 3.16, the full model is provided in Table B.2 in Appendix B). As described by Torres Cacoullos and Travis (2010), the behaviour of high-frequent individual lexemes clearly influences the outcome of the larger categories forming the factor group verb type.

Table 3.16: Comparison of log odds for verb type glm vs glmer

factor level	glm logodds	p-value	glm	glmer logodds	p-value	glmer
psychological	-0.667	0.0344	*	-0.467	0.2709	
auxiliary	-1.879	<.0001	***	-1.289	0.0051	**
modal	-0.920	0.0075	**	-0.859	0.0908	.

3.4.4 Overview results ICE Great Britain

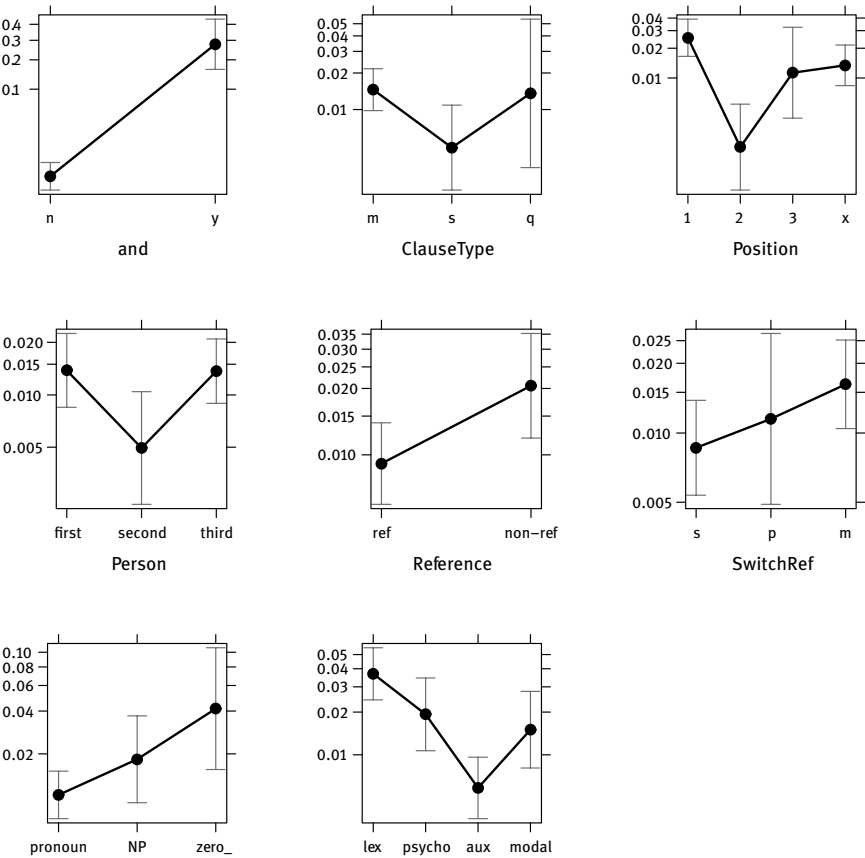


Figure 3.15: Predicted probabilities for factor levels ICE-GB

Figure 3.15 provides an overview of the effects of all factor groups in the minimal adequate model, showing the predicted probabilities (y-axis) of all factor levels (x-axis, the order is based on the order of factor levels in the model; for more details, see Fox 2016).

Mind that the scale of the graphs is adjusted to optimally display contrasts within each individual factor, so this graph does not accurately represent the weight of the different factor groups, but rather the relation and direction of effects of the different factor levels. The vertical lines represent confidence bands for each factor level, and are naturally more pronounced for contexts with lower token numbers (Fox 2016).

The following section presents a separate analysis for first person contexts. Interpretation and discussion of results for both models are provided in section 3.6.

3.5 Null subjects in ICE Great Britain: First person contexts

Two issues motivate a separate analysis for this subset of data: first person is the context for which relatively comparable studies on English exist; on the other hand, person has been presupposed as an influential factor a priori in preceding studies. Due to its comprehensive design, the present study is in the unique position to test this assumption.

3.5.1 Overview: Subject tokens and linguistic factors first person

The assumption that constraints on first person pronouns are different from those on second and third person is evaluated by testing the influence of the same structural factors on the more restricted context. To analyse first person contexts, a few modifications from the analysis of the full data set are necessary. Obviously, the factor groups person and reference are not relevant here. Moreover, since questions constitute an invariant context for first person pronouns (see also Torres Cacoullos and Travis 2014) the factor level is excluded from the factor group clause. This exclusion results in a total number of 901 overt and null subject tokens for analysis, with a null subject rate of 4.88%, analysed for eight possible structural determinants. Due to the rather low token numbers, the factor levels full and partial switch reference are merged, yielding the commonly employed binary distinction between continuous and switch reference. English null subjects in non-initial position are extremely rare, except in coreferential coordination, where the position is usually >3; the factor levels Position: 2 & 3 are

combined to cover non-initial, non-coordinated subject tokens. Table 3.17 provides an overview of the remaining categories and token numbers for first person pronouns for each factor level.

Table 3.17: Overview linguistic factors first person ICE-GB

		# zero	% zero	# overt	% overt	# total
subject		44	4.88	857	95.12	901
and	no	26	2.99	843	97.01	869
	yes	18	56.25	14	43.75	32
Clause type	main	43	5.62	722	94.38	765
	subordinate	1	0.74	135	99.26	136
Position	1	18	4.26	405	95.74	423
	2 & 3	4	1.68	234	98.32	238
	>3	22	9.17	218	90.83	240
Turn boundary	no	38	5.94	602	94.06	640
	yes	6	2.30	255	97.70	261
Turn length	1	2	2.78	70	97.22	72
	2	7	6.09	108	93.91	115
	3	6	4.55	126	95.45	132
	>3	29	4.98	553	95.02	582
Switch reference	maintenance	30	8.77	312	91.23	342
	switch	14	2.50	545	97.50	559
Persistence	pronoun	36	4.34	794	95.66	830
	NP	3	6.00	47	94.00	50
	zero	5	23.81	16	76.19	21
Verb type	lexical	17	11.89	126	88.11	143
	psychological	10	3.56	271	96.44	281
	auxiliary	11	3.27	325	96.73	336
	modal	6	4.26	135	95.74	141

3.5.2 Results first person

The model building and selection process follows the same procedure elaborated in sections 3.3.5, 3.4.1 and 3.4.2. Table 3.18 shows the Wald statistics indicating the statistical significance of the different factor groups for first person subjects.

Table 3.18: Wald statistics of factor group significance first person ICE-GB

Factor group	Chi-Square	d.f.	p
and	29.94	1	<.0001
Clause	4.11	2	0.0425
Position	5.00	3	0.1720
Turn boundary	1.64	1	0.1999
Turn length	4.97	3	0.1742
Switch reference	0.90	2	0.0012
Persistence	15.06	2	<.0001
Verb type	8.65	3	<.0001
Total	80.68	16	<.0001

As for the complete model, turn length and turn boundary are found not significant for subject pronoun realisation. In contrast to the full model, where switch reference is included with marginal significance, in the model for first person pronouns it bears no statistical significance at all. Surprisingly, position, one of the most significant factor groups for the full data set, is not significant here, potentially due to low token numbers in position 2 & 3 and the resulting merge of the two subcategories; this is even more likely given the stark decline of omission rates for second position in the full data set, and the following increase of null subject rates for position 3, and especially >3. Coordination and persistence remain highly influential factor groups, verb type and clause are kept in the model as well with clear statistical significance. Following the exclusion of the non-significant factor groups, Table 3.19 presents the minimal adequate model for first person null subjects in ICE-GB.

Despite the relatively low overall token numbers, thanks to the robust procedure the logistic regression model can be fitted to the first person subset with $p < .01$ in the Model Likelihood Ratio Test. Given the smaller data set, diagnostic values are necessarily lower than for the full model, with a concordance value $C = 0.83$, and 31% variation explained ($R^2 = 0.31$), which are still acceptable. Random resampling and bootstrapping largely confirm these diagnostics, with a corrected C of 0.81 and corrected R^2 of 0.26 (the full output is provided in Table B.14 in Appendix B).

Table 3.19: Logistic regression model first person ICE-GB

	Coefficient	SE	Z	p-Value	
Intercept	-2.6736	0.3657	-7.311	<.0001	***
and: y	3.4712	0.4426	7.842	<.0001	***
Clause: subordinate	-1.8672	1.0349	-1.804	0.0712	.
Persistence: NP	0.9470	0.6498	1.457	0.145	
Persistence: zero	2.4521	0.6710	3.654	0.0003	***
Verb type: psychological	-1.2887	0.5046	-2.554	0.0107	*
Verb type: auxiliary	-1.0269	0.4677	-2.196	0.0281	*
Verb type: modal	-0.8773	0.5604	-1.565	0.1175	

The variance inflation factors show no apparent interactions (Table 3.20).

Table 3.20: Estimated variance inflation factors first person ICE-GB

Factor level	vif
and: y	1.07
Clause: subordinate	1.02
Persistence: NP	1.06
Persistence: zero	1.11
Verb type: psychological	1.49
Verb type: auxiliary	1.43
Verb type: modal	1.33

Compared with the model for the full data set, first person lacks significance for switch reference and position. Significant factor groups for first person pronouns yield the following constraint ranking:

and > Persistence > Clause > Verb type

Table 3.21 provides a comparison of the significance of the different factor groups for the full and the reduced first person data set of ICE-GB conversations.

Table 3.21: Comparison of factor group significance all vs first person ICE-GB

Factor groups	Full set	Range	First person	Range
and	***	72	***	72
Clause	**	27	.	32

Table 3.21: (continued)

Factor groups	Full set	Range	First person	Range
Position	***	56	—	—
Turn boundary	—	—	—	—
Turn length	—	—	—	—
Switch reference		16	—	—
Persistence	**	35	***	53
Verb type	***	44	**	25

As in the full set, coordination strongly favours zero for first person. Subordinate clauses favour overt pronouns, although due to the low token numbers, this result is not very informative for the reduced set of first person contexts.

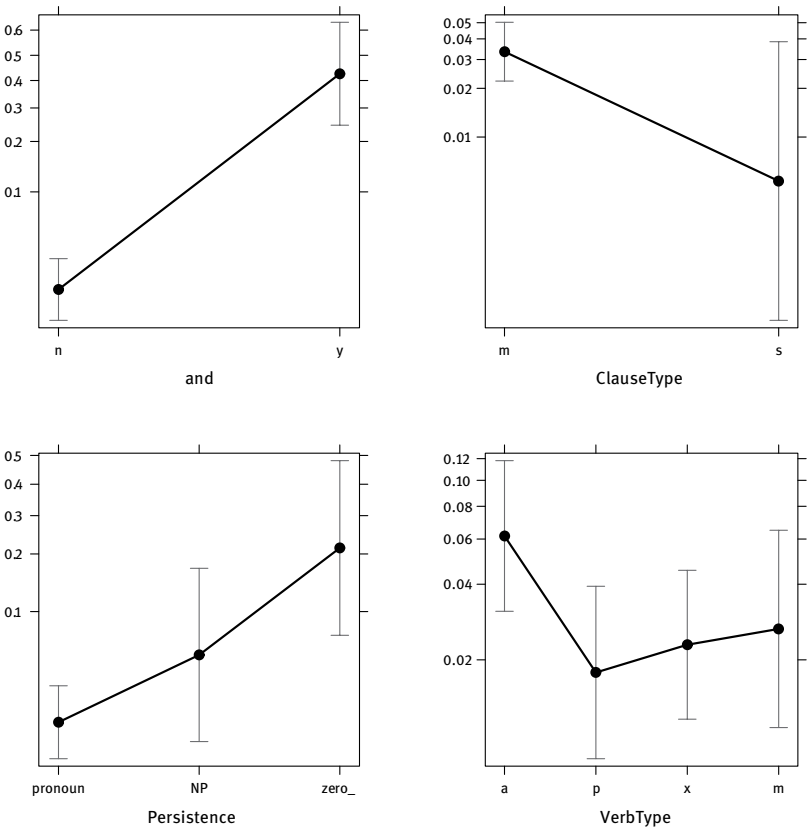


Figure 3.16: Predicted probabilities for factor levels first person ICE-GB

Persistence is even more important than in the full model, immediately preceding zero has a high omission triggering effect on first person pronouns. Verb phrase plays a smaller role concerning the strength of the factor group. The pronoun favouring effect of primary and modal auxiliaries is similar to the full model. A difference is found in the behaviour of psychological verbs: in combination with first person subjects their pronoun-preserving effect is indeed larger than for the full set, as predicted by e.g. Travis (2007).

Figure 3.16 provides an overview of the direction of effects in the first person model. The comparison to the full data set shows the same direction of effect at work for the factor groups coordination, clause type and persistence. Categorical differences between first vs second and third person pronouns thus cannot be confirmed for these factors, with the exception of questions as invariant context. On the other hand, the stronger pronoun-favouring effect for psychological verbs is confirmed for first person contexts; empty pronouns are even less likely for psychological verbs than for primary auxiliaries (see Figure 3.15). The vote is less clear for the statistically non-significant factor groups; their influence might just lie below measurable amounts, given the limited number of tokens included in this subset.

3.6 Summary and discussion

Quantitative investigation of null subjects in Standard English has been severely lacking until very recently. The present study joins the ranks of Torres Cacoullos and Travis (2014), Travis and Lindstrom (2016), and Wagner (2012, 2018) in the attempt to systematise the variationist analysis of this low-frequency syntactic phenomenon in standard and contact varieties of English, following the approach of Torres Cacoullos and Travis (2015) to contact varieties of Spanish. This section provides an evaluation of the results presented in the preceding sections 3.3, 3.4, and 3.5, including a comparison with previous studies on traditional L1 varieties of English. These results provide the basis for the comparative analysis of contact effects on English null subjects in chapters 5 and 6.

3.6.1 Discussion full data set

The present investigation of null subjects in informal spoken English shows that this non-standard grammatical phenomenon can meaningfully be described by means of a variationist analysis, in spite of its low frequency (as outlined in section 2.2). The observation of Torres Cacoullos and Travis (2014), and Travis

and Lindstrom (2016) that the envelope of variation is narrower for English is only partly confirmed: null subjects outside their predicted contexts of coordination, initial position, and declarative main clauses are found rarely, but they do occur (see section 3.1). The majority of the investigated 130 null pronouns are found in the predicted contexts, i.e. main clause (120), coordination (59), and utterance initial position (48), although none of these restrictions is categorical. These differences in attestation are partially afforded by the wider scope of the present study compared to earlier investigations, and underline the relevance of this more inclusive approach (section 3.2, see also section 2.3).

It is further confirmed that extra-linguistic demographic factors like speaker gender and age do not contribute significantly to the variable realisation of subject pronouns, confirming findings by Wagner (2012). The group of younger speakers shows more internal heterogeneity, and exhibits more variation towards higher deletion rates. In terms of overall null subject rates, the age groups are still very similar. However, the role of individual speakers for subject realisation rates is not negligible; almost one quarter of the speakers in the present data set produces invariant overt subject pronouns. Integrating speaker as a random effect is certainly a promising course of action, but requires higher token numbers to be computed reliably. The overall subject deletion rates found in this study are similar enough to those reported for spoken English in the literature to plausibly assume an accurate representation in the data here (see sections 3.3.1, 3.3.2, and 2.3.2). Confirming observations by Biber et al. (1999: 1105–1106), and Haegeman and Ihsane (2001: 333) on the higher tolerance of British English towards null subjects, the omission rates found here are apparently higher than those reported in Torres Cacoullós and Travis (2014: 22) for spoken American English, and closer to those found in Newfoundland English by Wagner (2012: 100–101), a variety described as particularly in favour of null subjects.

Collexemes and high-frequency verb tokens show an influence on subject realisation. Considering both verb lemmas and verb forms with unidirectional and bidirectional attraction measures, a number of verbs emerge as high attractors for null subjects in British English, especially negated auxiliaries like *doesn't*, *don't*, *can't* (section 3.3.3), and high-frequency tokens attracted to the coordinative construction like *go*, *get*, *come*, *see* for the first, and *say* for the second verb slot (section 3.4.3.1). This confirms findings by Teddman (2011) on collocational patterns in British English, and by Torres Cacoullós and Travis (2014) on “lexically specific” coordinations favouring null subjects as a cross-linguistic tendency, even beyond their investigated contexts of clause initial, and first person singular subjects, respectively.

The methodological issue of model evaluation is not resolved conclusively for logistic regression models (see e.g. Field et al. 2012, chapter 8), hence the report-

ing of model diagnostics is not as established in sociolinguistic studies as would be desirable. Still, to evaluate model fit it is not sufficient to consider the significance of factor levels only; diagnostic measures of logistic regression models, including overall model significance, the validity of results beyond the given data set (such as the concordance index, or validation via repeated random resampling), and a measure of the amount of variation explained (e.g. Nagelkerke's *R*) provide useful information on the reliability of the reported results (see e.g. Levshina 2015: chapter 10; see also section 3.4.2). All in all, the logistic regression model presented in Table 3.6 provides a satisfactory account of the role of linguistic factors for British English variable subject realisation, especially when compared to model diagnostics of earlier studies (provided that they are even available).

The distinction between three sources of factors, English-specific, contact, and universal, is mainly intended as an evaluation tool for the comparison conducted in chapters 5 and 6. For British English alone, these categories prove partly useful: The three major structural predictions for English null subjects are largely confirmed, as shown by the statistical significance of the three English-specific factor groups coordination, position, and clause in the multivariate analysis. In terms of weight of the factor groups, coordination bears the greatest statistical significance, followed by verb type and position. The statistical weight of clause is surprisingly low considering its status as an English-specific constraint and the apparent rarity of null subject tokens outside the context of declarative main clauses. Contact factors are of less statistical significance. Universal factors have diverging influence: factor-group significance is high for verb type, intermediate for persistence, and rather low, bordering on statistical insignificance, for switch reference. Other factors found to be influential in Wagner (2012), i.e. turn length and turn boundary, are not significant here. Overall, the English-specific factor groups provide the greatest explanatory value for the British English data set.

Coordination is the factor group with the strongest influence, but position and clause also behave as expected (section 3.4). More surprising is the relative relevance of the contact factors person, and specific reference. However, the comparative lack of second person zero contributing most strongly to the significance of the factor group person is probably not due to an underlying split-person system as found in partial null subject languages like Finnish, but rather based on communicative motivations and the role of second person as “genuine pronouns”, especially compared to the more varied functions of third person pronouns, and the apparent lower ambiguity of first person reference by physically present speakers in real-time, face to face conversation. The importance of overt specific second person reference in conversation is also attested for other languages (see sections 2.2 and 4.1), and corroborated by the following observation:

of the rare second person null pronouns, half represent non-specific, generic reference. The other source of non-referential null pronouns is based in large parts on utterance initial expressions like *depends*, *seems*, *sounds*, *looks* (see also Teddiman 2011). However, while it is present, the favouring effect of non-referentiality for zero expression is small, and statistically barely significant.

The cross-linguistically widespread factor switch reference is not as relevant for English as for other languages, in spite of a slight favouring effect of reference maintenance. This confirms findings by both Torres Cacoulllos and Travis (2014) and Wagner (2012, 2018). Clear effects are measured for the other universal factor groups persistence and verb type. Preceding zero has a strong triggering effect for further zero subjects, especially for first person pronouns. Concerning different verb types, null subjects are most likely for lexical verbs, while primary and modal auxiliaries clearly favour pronouns. This seems to go against Wagner's (2012, 2018 especially) relation of the occurrence of null subjects with increasing VP complexity, which is mainly achieved through auxiliaries. However, verb type as a factor group provides better explanatory value than VP complexity for the present data set (see Table B.9 and Table B.10 in Appendix B).

3.6.2 Discussion first person

Differences of first person pronouns to the full set are addressed in section 3.5; this section is devoted to a comparison with earlier studies on first person null subjects in L1 varieties of English (Torres Cacoulllos and Travis 2014, Wagner 2012, 2018). While there is no entirely parallel comparison possible due to different study and variable designs, an explicit comparison with the results of the three major preceding corpus studies on English is attempted. Table 3.22 provides an overview of the shared factors in the four studies.

The first crucial difference between the studies compared lies in the circumscription of the variable context. Coordinated contexts are not part of Wagner's analyses. Given the central role of this factor in the other studies, this fact makes the omission rates for Newfoundland English even more remarkable, and confirms descriptions of the variety as rather liberal concerning subject omission. A further difference is found regarding the clause types analysed. While Wagner (2012) includes the rare null tokens in subordinate clauses, they are not part of her follow-up study on the same data (2018), and not included in Torres Cacoulllos and Travis (2014). Contexts outside declarative main clauses are indeed marginal in the present study, especially in the subset of first person compared to the full data set; as argued in sections 3.1 and 5.1, their inclusion is not least for the benefit of the Asian varieties analysed in the following chapters.

Table 3.22: Comparison of factor group significance first person contexts ICE-GB, US English, and Nfdl English

	Wagner (2012)	Wagner (2018)	Torres C. & Travis (2014) ¹	ICE-GB first person
variety	Nfdl	Nfdl	US	British
tokens analysed	8,514	4,025	302 [6,700]	901
% zero	4.4%	7.2%	50% [2.2%]	4.9%
Coordination	NA	NA	***	***
Clause	*	NA	NA	.
Position	.	**2	**	—
Turn boundary	—	NA	—	—
Turn length	**	**	NA	—
Switch reference	—	.	—	—
Persistence	***	***	**	***
Verb type	.	*	—	**
VP complexity	**	**	— ³	—

NA = factor not included in the respective study, — no statistical significance, ./*/**/*** indicating increasing significance, see also Table 3.6

¹ The 151 null subject tokens analysed by Torres Cacoullos and Travis (2014) are based on texts containing a total of 6,700 subjects; the multivariate analysis is based on a 1:1 ratio of 151 tokens each for overt and null pronouns

² subject slot

³ polarity

The only factor group that is identified as a highly significant predictor in all four studies is persistence. An interesting observation can be made regarding the nature of this factor group: it seems indeed to be independent of referential continuity. Torres Cacoullos and Travis (2014) model persistence as preceding null coreferential pronouns, Wagner (2012, 2018) and the present study include preceding null tokens regardless of reference. Still both types significantly influence subject omission. On the other hand, switch reference, indicating a reference maintaining function of null pronouns, rather than a more mechanical “priming effect”, is rendered insignificant in three of the studies, and barely significant in Wagner (2018). In spite of the conceptual problems of zero as a triggering (non-) element, this persistence effect is apparently robust across L1 varieties of English.

Concerning the influence of verb type, insights are limited by the relatively low token numbers – studies on lexically determined variation in natural languages are notoriously in need of large volumes of data. It is likely that the lack of effect for this category in Torres Cacoullos and Travis (2014) is based on their

selective sampling method, which excludes the large majority of overt pronominal tokens.

Verb phrase complexity, on the other hand, shows significance only for Wagner's Newfoundland data. A simplified version is present in Torres Cacoullos and Travis' factor polarity, distinguishing between negated and non-negated verb forms (negation constitutes one of the "sense units" increasing VP complexity, Wagner 2018: 16). In the present study, the factor verb phrase complexity is conceptualised following Wagner's model, but does not show a statistically significant effect, especially when compared to the factor verb type. One point to consider regarding this difference in significance concerns the kinds of data analysed in the present study compared to Wagner (2012, 2018). Wagner's corpus of Newfoundland English consists of interviews "[following] the traditional oral-history type/sociolinguistic interview method" (Wagner 2012: 100). Topics include family and local history, and the discussion of local stories, customs, and personal experiences, commonly with reference to past events. This is possibly represented by a relatively high amount of more complex verb phrases in Wagner's data (VP1 = 30%, VP2 = 50%, VP3+ = 20%) compared to the present data set (VP1 = 43%, VP2 = 44%, VP3+ = 13%), which consists largely of naturally occurring conversation. The influence of text type on subject realisation is also addressed in section 6.2.

Overall, the contrastive view on the four studies shows both the potential and the need for further empirical analysis of English null subjects based on analytical conventions that are only just emerging.

3.6.3 Conclusion

The preceding analysis has evaluated the status of common assumptions about English null subjects. Travis and Lindstrom's (2016: 104) claim that "English is not special at all" with regards to null subjects cannot be affirmed enthusiastically. Null subjects in British English are most strongly determined by language specific and partly by universal constraints, while the influence of contact features is more limited.

The comparison with previous research stresses the importance of transparent study design: different studies define the envelope of variation slightly, or even radically differently, their results are then not directly comparable. Exclusions from the variable context are not always open to scrutiny, which makes it difficult to evaluate how justified they are. On the other hand, a study as inclusive in structural contexts as the present one deals with a rather mixed bag of cases and requires meticulous coding and thoughtful assessment of possible interactions to tease apart the decisive factors from epiphenomena.

The results achieved in this chapter do not offer great surprises concerning the linguistic predictors, but fill the research gap of sound quantitative investigation of null subjects in Standard British English, and serve as a solid basis for the evaluation of these structural constraints in Asian varieties of English in the following chapters.

Part B: **Null subjects in Asian Englishes**

4 Asian languages and varieties of English: Theory, description and comparison

[Those] instances of deviation from the norms of either language which occur in the speech of bilinguals as a result of their familiarity with more than one language, i.e. as a result of language contact, will be referred to as *interference* phenomena. It is these phenomena of speech, and their impact on the norms of either language exposed to contact, that invite the interest of the linguist. [...] Great or small, the differences and similarities between the languages in contact must be exhaustively stated [...] as a prerequisite to an analysis of interference. (Weinreich 1963: 1–2, emphasis in the original)

For the Asian contact varieties of English considerations about fundamentally different language types play a crucial role, since they are influenced by languages of a typologically different configuration with respect to subject pronouns. While English is often discussed as a prime example of a non-null subject language with a syntactic orientation (see section 2.3.1), Chinese is commonly presented as the prototype of a pragmatic or discourse-oriented language with ubiquitous null subjects. To fulfil Weinreich's request cited above, the first part of the present chapter is devoted to a detailed description of null subjects and their conditions in these analytic NSLs.

Section 4.1 examines the role of subject pronouns in this language type, which is especially pervasive in Asia, covering formalist (section 4.1.1), functionalist (section 4.1.2) and variationist (section 4.1.3) approaches. Chinese, the best described of these languages, serves as the representative.

Section 4.2 is concerned with processes of language contact and their outcome (section 4.2.1), the role of subject omission in learner and contact languages (section 4.2.2), English in contact (*World Englishes*, section 4.2.3), and null subjects in contact varieties of English (section 4.2.4).

The Asian Englishes under investigation here are introduced in section 4.3 (India, section 4.3.1, Hong Kong, section 4.3.2, Singapore, section 4.3.3). Each subsection includes a sketch of the socio-historical background of their genesis, their linguistic ecology, i.e. the local substrates, and the traces they left in the morphosyntax of the respective variety. The structural descriptions also provide an overview of previous studies on the grammar and, as far as available, null subjects in the individual varieties.

The last section 4.4 is devoted to the productive field of comparative studies in World Englishes, arguing for the varieties in the present study as a meaningful unit of comparison (section 4.4.1) and clarifying the methodological principles for the comparative analysis in chapter 5 (section 4.4.2).

4.1 Null subjects in Asian languages

Parallel to the description of canonical null subject and non-null subject languages (NSL) in chapter 2, this section presents an overview of the conceptualisation of null subjects in non-canonical, or “radical pro-drop” languages (section 4.1.1), its analysis as the wider reaching syntactic configuration of *topic prominence* (section 4.1.2), followed by a discussion of relevant empirical research (section 4.1.3). Illustrative examples are mainly taken from reference grammars of (Mandarin) Chinese, the best described language of this type.

4.1.1 Radical pro-drop

As introduced in section 2.1, the term *radical pro-drop* was coined to describe analytic languages that use variable subject pronouns without the licensing effect of verbal morphology (C.-T. Huang 1984). Judging from the 711 languages sampled in the *World Atlas of Language Structures* (WALS), optional and obligatory subject pronouns are similarly common cross-linguistically. Both types are much rarer overall than subject affixes, which imply canonical pro-drop³; overt subject pronouns can thus be considered the typologically marked case. The feature shows a clear areal distribution: obligatory subject pronouns, i.e. non-NSLs, are predominantly found in (Northern) Europe (see also Haspelmath 2001) and West Africa, while optional subject pronouns without affixation, or radical pro-drop languages, are especially common in Australo-Pacific and East and Southeast Asian languages, including Sinitic languages (Dryer 2013).

In stark contrast to English as the paragon of non-pro-drop, and Spanish, Portuguese and Italian as the most cited canonical pro-drop languages, Chinese is usually discussed as the prototype of non-inflecting pro-drop (D'Alessandro 2015: 203). Concerning the denotation of Chinese, Matthews provides the following definition:

In a general sense, *Chinese* covers any of the languages and dialects belonging to the Sinitic language family and standing in a certain relationship to the logographic Chinese script. [...] Commonly, unless otherwise specified, ‘Chinese’ refers to the standard language, often without distinction between written and spoken forms. The term ‘Mandarin’ is used either to refer to the standard language, or to delineate a group of dialects spoken natively in northern and western parts of China. (Matthews 2010: 757)

³ WALS feature 101A: 82 languages have obligatory pronouns, 61 optional pronouns, 437 use affixes, see Dryer 2013.

Both *Chinese* and *Mandarin* will be used in this sense, unless specified otherwise. Differences between varieties of Chinese are subtle in the domain of morphosyntax and discourse structure; Chao even claims that “there is practically one universal Chinese grammar” (Chao 1968: 13). It is therefore common, and justified by the wide-reaching structural similarities, to use Mandarin Chinese as a reference variety for discussing grammatical phenomena, like the extent and conditions of null subjects (see e.g. Ansaldo 2010: 514; Bao 2012: 167, Matthews and Yip 1994: 5). Examples from reference grammars of Chinese in the following sections are provided in Pinyin romanisation.

Zero anaphora is but one of several related features of these languages, which C.-T. Huang describes as having “maximal freedom for the use of zero pronouns” (C.-T. Huang 1984: 534). Like canonical NSLs, radical NSLs do not use expletive subjects (see also section 2.1) for existentials (4.1), extraposition (4.2), or raising (4.3). Instead of pleonastic pronouns, for weather predicates a full thematic noun phrase (*yǔ* ‘rain’) is obligatory (4.4).

- (4.1) *Yǒu yī-ben shù zài zhuōzi shàng.*
have one-CL book on table top

‘(There) is a book on the table.’

(Chu 1998: 57)

- (4.2) *Shì Zhāngsān dà le Lǐs.*
is Zhangsan beat LE (CRS) Lisi

‘(It) is Zhangsan (who) beat Lisi.’

(Xiao 2002: 237)

- (4.3) *Kànshangqu Zhāngsān hěn lèi.*
seem Zhangsan very tired

‘(It) seems (that) Zhangsan is very tired.’

(Yuan 1997: 473)

- (4.4) *Xià yǔ le.*
descend rain LE (CRS)

‘(It’s) raining.’

(Li and Thompson 1981: 91)

Non-specific subjects can be encoded by generic full noun phrases (such as *rén* ‘people’ or *yiban rén* ‘people in general’), but strongly favour zero expression (4.5).

- (4.5) *Yóuyǒng yǐqián,* .
 swim before
yìbān yào xiān zuò zhǔnbèi yùndòng.
 generally ought first do preparation exercise
 'Before (one) swims, (one) generally must do warm-up exercises.'
 (Chu 1998: 288)

For referential pronouns, zero anaphora is permitted not only for subjects (4.6 b, d, e, f), but is also widespread for objects (c - f), both in main (b - d) and embedded clauses (e - f).

- (4.6) **Speaker A / Question:** .
Zhāngsān kànjian Lǐsi le ma?
 Zhangsan see Lisi LE Q
 'Did Zhangsan see Lisi?'

Speaker B / Answer:

- a. *Tā kànjian tā le.*
 He see he LE
 'He saw him.'
- b. *Ø Kànjian tā le.*
 '[He] saw him.'
- c. *Tā kànjian Ø le.*
 'He saw [him].'
- d. *Ø Kànjian Ø le.*
 '[He] saw [him].'
- e. *Wǒ cāi Ø kànjian Ø le*
 I guess see LE
 'I guess [he] saw [him].'
- f. *Zhāngsān shuō [Ø kànjian Ø le].*
 Zhangsan say see LE
 'Zhangsan said that [he] saw [him].'

(C.-T. Huang 1984: 533)

Variant a. demonstrates the occurrence of both overt subject and object pronouns in Chinese. Although zero anaphora is considered the norm, and overt pronouns are described as the marked case (Li and Thompson 1979: 327), null subject rates in Chinese hover around 50% of all pronominal subjects (compared to up to 80% zero in canonical NSLs like Spanish, Torres Cacoullós and Travis 2014: 22). None of the variants (b – f) is acceptable in a non-NSL; null objects are not regularly part of canonical NSLs either. Reference in radical NSLs is thus consistently less explicit; identifying referents demands a greater amount of interpretive work from the hearer or reader. C.-T. Huang (1984: 531) refers to this different degree of audience participation as *cool* (i.e. content is left implicit, “much audience participation required”) vs *hot* languages (i.e. content is coded explicitly, “little audience participation required”).

Syntactic constraints on zero pronouns in Chinese are only attested for the contexts following co-verbs (*gēn* ‘with’, 4.7), and as pivotal noun phrase in a serial verb construction (should be *nǐ* ‘you’, 4.8).

- | | | | | | | |
|-------|-----|------|---|-------|----------|------|
| (4.7) | *wǒ | gēn | Ø | xué | Yīngwén. | |
| | I | with | — | learn | English | |
| | | | | | | |
| (4.8) | *wǒ | quàn | Ø | bié | hē | jiǔ. |
| | I | urge | — | not | drink | wine |

(Li and Thompson 1981: 675)

Zero anaphora in radical NSLs is licensed on discourse rather than sentence level, reflected in the alternative designation *discourse pro-drop*. Identification is granted via the ability of “zero-topic” to bind empty variables like null pronouns (Huang 1984). In non-generative approaches, radical NSLs like Chinese have alternatively been described as *pragmatic* or *discourse oriented* (vs *syntactic* or *sentence oriented*, Tsao 1977, see Givón 1979a for a parallel classification of different “modes of communication”).

Chu (1998: 1) remarks on the somewhat vague definition of Chinese sentences as “expressing a complete thought”, and “having a predominantly *liúshuǐjù* (literally, ‘flowing-water sentence’) structure”. However poetic, vague or impressionistic, these observations are common in academic descriptions of Chinese. An alternative widely employed classification of languages based on similar characterisations, stressing discourse over syntactic factors, has been proposed by Li and Thompson (1976) and is discussed in the next section.

4.1.2 Topic prominence

The universal status of the grammatical subject-predicate dichotomy has been challenged by typological research, which reveals the cross-linguistic frequency of a primarily discourse-semantic based configurationality. Besides the obligatory linguistic expression of subjects in canonical declarative clauses, this affects reflexivisation, passivisation, verb serialisation and imperatives (Kiss 2001: 1442). Chao (1968) first applies the semantic notions of *topic* and *comment* to Chinese sentence structure. He determines them as the primary constituents of basic sentence structure in Chinese, an interpretation that has been widely accepted ever since (see e.g. Xu 2015: 394). Topic and subject of a sentence usually coincide. In the absence of other topicalised elements the subject is the topic of the sentence by default; subjects are sometimes even described as “grammaticalised topics” (Kibrik 2001: 1414). Disagreement occurs regarding the question of universality of subjects as a sentence constituent. Generative analysis of null subjects in languages like Chinese assumes pro-drop via recoverability of the referent in discourse (rather than verbal phi-features, see Camacho 2013: 125–139 for a summary). In contrast, Li and Thompson (1976) challenge the universal status of the syntactic category subject. They argue for a “basic” topic-comment structure of sentences in topic-prominent languages like Japanese, Korean, and Chinese, in contrast with subject-prominent languages from the Niger-Congo and Indo-European language families, including English (Li and Thompson 1976: 460). Subject-predicate structures in topic-prominent languages are then just a specific, non-obligatory subtype of the basic topic-comment structure.⁴

Topic is defined as “what the sentence is about”, on a larger scope, it sets “a spatial, temporal, or individual framework within the main predication holds” (Chafe 1976: 50). Although topic is employed as a syntactic notion for topic-prominent languages, its identification requires the discourse context (Chu 1998: 250). Li and Thompson’s (1976) seminal account of topic-prominent languages continues its formative influence on the description of non-inflectional NSL to this day. According to Li and Thompson (1976: 461–465), topics differ from subjects in the following respects:

- a. They must be definite;
- b. They need not be an argument of the predicative constituent (“selectional relations”);

⁴ Recent generative approaches have attempted to reconcile topic prominence and pro-drop, see e.g. Modesto (2008), who argues for pro-drop as a sub-parameter of the macro-parameter topic prominence.

- c. The verb determines the subject, but not the topic (i.e. action verbs require “actor” as subjects, stative verbs “patient/experiencer”, etc.);
- d. Topics carry a specific functional role (topics are the centre of attention, and constant across sentences);
- e. They need not show verb agreement;
- f. They must be in sentence-initial position;
- g. They have no central role in grammatical processes like reflexivisation, passivisation, etc.

Some of these properties are related, e.g. analytic languages like Mandarin commonly employ word order to mark definiteness; initial position in this case constitutes both topic- and definiteness-marking functions at once (Li and Thompson 1981: 23). Topic-prominent languages permit sentences with both topic (*nèi-zhī gǒu*, ‘that dog’) and subject (*wǒ* ‘I’, 4.9), coincidental topic and subject (*wǒ* ‘I’, 4.10), but also with topic (*zhèi-ge tí-mù* ‘this topic’) and without overt subject, e.g. when the subject is implied as non-specific referent (4.11).

- (4.9) *Nèi-zhī gǒu wǒ yǐjīng kàn-guo le.*
 That-CL dog I already see-EXP LE (CRS)

‘That dog I have already seen.’

(Li and Thompson 1981: 88)

- (4.10) *Wǒ xǐhuān chī píngguǒ.*
 I like eat apple

‘I like to eat apples.’

(Li and Thompson 1981: 88)

- (4.11) *Zhèi-ge tí-mù zài hǎo buyǎo tí chū lái.*
 This-CL topic most good don’t bring:up exit come

‘[This topic]_i, (you’d) better not bring it_i up.’

(Li and Thompson 1981: 89)

While the topic-comment structure constitutes the basic sentence pattern in topic-prominent languages, English usually employs syntactic movement to derive marked structures like prepositional phrases (4.12), left dislocation (4.11), and preposing, or topicalisation (4.13).

- (4.12) Concerning / speaking of / as for [that dog]_i, I’ve already seen it_i.

- (4.13) Apples_i I like to eat Ø_i.

Chinese-style topics on the other hand have no direct counterparts in English (Chafe 1976: 50). Lambrecht (1994: 30) describes the difference as one between linked and unlinked topic constructions. Chinese-style topics are usually related to the comment in a part-whole or set-subset relation (a few volumes out of a set of books, 4.14), whereas English topics are analysed as anaphorically linked to an overt (4.12) or empty (4.13) pronoun.

- (4.14) Zhè tào shū, nǐ kěyǐ ná zǒu jǐ běn.
this set books you may take away a-few volumes

‘You may take away a few volumes of this set of books.’

(Xu 2015: 395)

Topic-prominent languages are more common cross-linguistically than subject-prominent languages (e.g. Kiss 2001: 1442). Topic prominence, or the “pragmatic mode” is also a feature of early child as well as adult learner interlanguage (Givón 1979b: 226–227, Fuller and Gundel 1987). Li and Thompson (1976: 466–471) list the following properties for topic-prominent languages:

- a. Surface coding of topic, but not necessarily subject (by overt topic markers in Japanese, or initial position in Chinese);
- b. Passivisation as either a marginal construction, or absent altogether;
- c. Absence of dummy subjects;
- d. Double subjects (4.17);
- e. Control of co-referential constituent deletion by topic (4.15);
- f. Tendency for verb-final clause structure;
- g. Absence of constraints on topic constituents (4.16);
- h. Basicness of topic-comment sentences.

Especially relevant here are features a. and c., which imply the possibility of null subjects. Testable contexts within the scope of the present study are also presented by properties e. (4.15) and g. (4.16, “time phrase as topic”).

- (4.15) Niè kuài tiān .
that piece land

dàozu zhǎngde hěn dà, suǒyǐ Ø hěn zhìqián.
rice grow very big so Ø very valuable

‘[That piece of land]_{TOPIC}, rice grows very big, so it (the land) is very valuable.’

(Li and Thompson 1976: 469)

- (4.16) *Zuótiān xuě xià de hěn jǐn.*
 yesterday snow descend CSC very incessant

‘Yesterday_{TOPIC} it snowed incessantly.’

(Li and Thompson 1981: 94)

According to Li and Thompson (1976: 25–26), *double subjects* are pervasive in topic-prominent languages. Similar to Chinese-style topics, the two noun phrases (*xiàng* ‘elephant’, *bízi* ‘noses’) are usually related by a possession or subset relation (4.17).

- (4.17) *Xiàng bízi cháng.*
 Elephant nose long

‘Elephants’ noses are long / Elephants have long noses.’

(Li and Thompson 1981: 92)

Addressing the issue of double subjects, Chu (1998: 30) analyses the layered topic-comment structure of complex sentences (square brackets indicating the first, round brackets the second layer; adapted from Chu 1998: 29).

- (4.18) *Zhè sānshí-ge xuésheng,*
 this 30-CL student,
èrshíbā-ge bù jígé yīdìng děi bùkǎo.
 28-CL not pass Ø definitely Ø must make-
 up-exam

‘Of these thirty students,

twenty-eight failed (the test) (they) must take a make-up exam.’

[_{top} zhè 30ge xuésheng] [_{com} (_{top} 28ge) (_{com} bù jígé) (_{com} yīdìng děi bùkǎo)]

Chu terms the designation “double subject” a “misnomer”; the first NP *zhè sānshí-ge xuésheng* (‘these thirty students’) is unambiguously analysed as topic, while the second NP *èrshíbā-ge* (‘twenty-eight ones’) can be analysed as either subject or topic, embedded within the comment. This layered structure, can – in theory – be infinite, and resembles embedding in the syntactic structure of a sentence. Equally, the actual depth of layering is limited by speakers’ cognitive capacities. The descriptive term *double NP* is thus preferable for these structures (Chu 1998: 31).

A common discourse function of explicit referent introduction in Chinese is to build the point of departure for a *topic chain* (Li and Thompson 1981: 659). Following its establishment via an existential construction, in accordance with Givón's "quantity principle" (see section 2.1.2), a stable referent can in subsequent discourse repeatedly be referred to by zero anaphora (4.19).

(4.19)	<i>Luòyáng</i>	<i>yǒu</i>	<i>ge</i>	<i>míng</i>	<i>gēnǚ_i</i> ,			
	Luoyang	had	CL	famous	song-girl			
	\emptyset_i	<i>jiào</i>	<i>Yáng</i>	<i>Zhuluo</i> ,				
		called	Yang	Zhuluo				
	\emptyset_i	<i>cōnghuì</i>	<i>guò</i>	<i>rén</i> ,				
		intelligent	surpass	people				
	\emptyset_i	<i>yǐ</i>	<i>yǔyán</i>	<i>jiānqiǎo</i>	<i>guan</i>	<i>yú</i>	<i>yīshí</i> .	
		with	language	sharp-skillful	top	at	the-time	

'In Luoyang, there was a famous female singer by the name of Yang Zhuluo. She was extremely intelligent and was tops for her quick wit.'
(Chu 1998: 261)

The prerequisite for topic chains is a high degree of conjoinability between clauses, provided by the relative proximity to the antecedent *gēnǚ* ('singer') and the absence of competing referents (Li and Thompson 1979: 320). In contrast, this conjoinability is disturbed by interruptions like topic switch, the insertion of adverbials, or change from background to foreground information and vice versa (Y. Huang 2000: 327). (4.20) illustrates this switch, where a series of descriptive background clauses forms the topic chain.

The lexical noun phrase in line 1 establishes the topical referent, the following topic chain is connected by zero anaphora, until the switch to foregrounded action triggers an overt pronoun in line 4.

- (4.20) 1 *Bái Xiānsheng zài kètīng-lǐ děng Lǐsì,*
 Bai Mr at living wait Lisi
 room-in
- 2 *Ø dài-zhe yǎnjìng, zài nàr kàn bàozhǐ,*
 wear-DUR glasses at there read newspaper
- 3 *Ø hǎoxiàng yǒu diǎn bu-nàifan,*
 seem have a:little not-patient
- 4 *tā shuō: "..."*
 he say

'Mr Bai was waiting for Lisi in the living room. (He) was wearing glasses and reading a newspaper there. (He) seemed to be a bit impatient.

He said: "..."

(Li and Thompson 1981: 663)

Given Gundel et al.'s (1993) observation on the frequency of referential expressions on the opposite ends of the givenness scale, in languages permitting empty categories zero anaphora is the default referential expression for given information (see section 2.1.2). As the marked case, overt pronouns can therefore be interpreted as contrastive focus markers, or as "highlight[ing] the referent of the pronoun in the context in which it occurs" (Li and Thompson 1981: 674). This is especially needed when semantically compatible referents are mentioned in the immediately preceding discourse, leading to "potential referential interference" (Givón 1983: 11).

Null pronouns in Chinese occur freely and are not restricted syntactically. The restriction to declarative main clauses found for English is not attested, neither is the constraint to initial position. No account of a principled preference for specific persons or verb types can be found, while persistence effects are likely due to the widespread use of topic chains. The absence of expletive pronouns in topic-prominent languages, and their lack of semantic value in English makes them a possible target for transfer. The same goes for generic pronouns, whose contribution to sentence meaning is negligible. The use of overt pronouns largely for contrastive marking might lead to a high usage in possibly ambiguous contexts, such as partial switch reference. Evidence for this is also found in the empirical studies on spoken Chinese discussed in the next section.

4.1.3 Quantitative research on Chinese null subjects

Quantitative analysis of null subjects in analytic NSLs is hard to come by; most accounts maximally provide rough estimates of subject omission rates, without further explanation or evidence. This section discusses two sociolinguistic studies on Mandarin Chinese conversational data (Jia and Bayley 2002, Li et al. 2012) that test conditions on subject omission found in canonical NSLs like Spanish for spoken Mandarin Chinese. Both involve a multivariate quantitative analysis comparable to the analysis in chapter 3.

Jia and Bayley (2002) analyse the speech of adult educated native speakers of Mandarin, consisting of telephone conversations among family members and friends, and teacher speech in Chinese language heritage schools in the US. They investigate the realisation of subject pronouns with human referents, and test the influence of factors such as discourse context, sentence type, subject person and number, and switch reference. The rate of null subjects in variable contexts, i.e. excluding co-verbs, serial verbs and invariant formulaic expressions such as greetings, is 47% zero of 1,400 overall subject tokens (Jia and Bayley 2002: 104). Jia and Bayley find that pronoun variation is systematically conditioned by multiple linguistic constraints. More specifically, overt pronouns are favoured by declarative sentences compared to questions and imperatives, switch reference, and first person singular. The latter two favouring conditions are widely attested in studies on inflectional languages as well. The overall distribution of null subjects, especially the factor “person and number”, crucially depends on the discourse context: it is found that overt pronouns are favoured by referents “unexpected” in the specific discourse situation, i.e. second person singular referents in classroom, where the teacher is more likely to address the group as a whole, and second person plural referents in conversational speech, where the dialogic setting of telephone conversations makes second person singular reference the more predictable option (Jia and Bayley 2002: 110–111).

Li et al. (2012) offer the most comprehensive multivariate analysis of Chinese null subjects to date. They analyse both animate and inanimate subject pronouns in free speech, including the Chinese *Pear Stories*, and classroom speech of both teachers and pupils. Factors include four social (age, occupation, gender, discourse context) and four linguistic categories (subject person and number, co-reference with subject of the preceding clause, specificity of subject referent, i.e. specific vs generic, and sentence type, Li et al. 2012: 92). Their overall null subject rate is 52.8% of 8,500 subject tokens. Overall, linguistic factors clearly outrank social factors. Most decisive are the factor switch reference, with partial overlap and complete switch most strongly favouring overt pronouns, and the factor person and number, with plural verbs favouring zero (Li et al. 2012: 103–104). As

expected, there is a favouring effect of non-specific reference for zero, although it is smaller than expected (Li et al. 2012: 112).

Both studies succeed in showing that null subjects in Chinese are systematic, and conditioned by factors also found for canonical NSLs; their findings match on the favouring zero effect of singular reference, declarative clauses, and reference continuity. Switch reference is coded as a three-way rather than a binary factor in both studies, with significant results for Chinese subject realisation: partial switch tends to pattern with full switch, or trigger even higher overt subject realisation rates (Jia and Bayley 2002: 108, Li et al. 2012: 104). Both studies further prove the importance of taking into account the discourse context, especially concerning differences between persons with regard to contrasts between conversations and classroom speech. However, there are no marked differences in the frequency of null subjects between narratives and conversations. While the two studies are not completely equivalent to the following analysis, they provide additional significant factors to those discussed in section 2.3 on English null subjects, and thus point towards possible “conflict sites” worth testing, such as the referential status of the pronoun, person, and switch reference.

The preceding sections have shown how radically different Asian languages behave compared to English with regard to subject pronouns. In order to evaluate the effect of typologically different substrates on the Asian contact varieties of English, the next section discusses key principles of language contact, before looking at the manifestation of these principles for the feature and varieties under investigation here.

4.2 Language contact and varieties of English

No language is an island; there is no living language entirely in seclusion. It is therefore not surprising that the structural and social impact of language contact has inspired a wealth of research, from Hugo Schuchardt in the late 19th century onwards. General principles of language contact and influential approaches to contact linguistics are sketched in section 4.2.1. The null subject parameter, as introduced in chapter 2 and extended to analytic languages in section 4.1, has received considerable attention with regards to its development in language contact, both in the context of second language acquisition, as well as its manifestation in creole languages (section 4.2.2). English as a global language has come into contact with diverse language types. Section 4.2.3 introduces frameworks of studying English in contact situations and its outcomes, the so-called *World Englishes*; followed by an account of null subjects in (other) varieties of English and English-lexifier creoles (section 4.2.4).

4.2.1 Language contact and contact languages

One point of contention in the investigation of language contact has been the *locus of contact* – is it primarily the bilingual speaker, as Weinreich (1963: 1) states, or is the interaction of two grammatical systems the central investigative field for linguists (Matras 2009: 2)? In the end, it is a collection of linguistic features that usually forms the backbone of language description, but it is obvious that, besides general principles of language change and the typology of the languages involved (e.g. Heine and Kuteva 2005: 37), research on language contact needs to take into account the socio-historical circumstances of contact (e.g. Thomason and Kaufman 1988, Mufwene 2001) and issues of acquisition and learning (e.g. Winford 2003: chapter 7). The role of these aspects is discussed in turn.

Refuting earlier assumptions by e.g. Antoine Meillet, Edward Sapir, or Roman Jakobson (see Thomason 2001b: 63), decades of research on the processes and results of language contact have shown that “essentially any part of language structure can be transferred from one language to another” (Heine and Kuteva 2005: 1). Depending on the research focus and structural level, this process has been termed *transfer*, *interference*, *borrowing*, *calque*, or *replication* (Lange 2012: 33–43). The most neutral term “transfer” is used here, with specified exceptions where necessary.

Heine and Kuteva are specifically interested in the historical process of contact-induced grammaticalisation via *grammatical replication* from the *model language* (Heine and Kuteva 2005: 2). This replication describes not a mere copying of a grammatical category, but rather initiates creative adaption and development of innovative use patterns in the target (or: *replica*) language (Heine and Kuteva 2005: 37). An example is the use of possession verbs as existential markers, illustrated by the Singlish *got*-construction. In a cross-linguistically common process, the Chinese possession verb *yǒu* (‘have’, 4.21) developed into an existential marker (4.22).

- (4.21) *Wǒ yǒu yī-běn xīn-de shū.*
I have one-CL new-DE book

‘I have a new book’
(Chu 1998: 55)

- (4.22) *Ø Yǒu yī-běn shū zài zhuōzi shàng.*
have one-CL book on table top

‘There is a book on the table’
(Chu 1998: 57)

Singlish speakers then “rely on the model provided by their L1 Chinese, thereby replicating the grammaticalisation process from possession verb to existential marker” (Heine and Kuteva 2005: 93; 4.23, see Lee et al. 2009 for a more detailed discussion of the origins and usage patterns of *got* in Singlish).

- (4.23) Here Ø got many nice houses.
 ‘There are many nice houses here.’
 (Lee et al. 2009: 295)

In their analyses of the multilingual ecology in Singapore (see also section 4.3.3), both Ansaldo (2009) and Bao (2010) emphasise the importance of rigorous assessment of the substrate languages in language contact situations. Bao’s (2010, 2015) usage-based approach identifies certain instances of transfer, as observed e.g. in Singlish, as *systemic*, in the sense that an entire grammatical subsystem is affected. A crucial component in systemic transfer is the convergence of the substrate languages involved, as shown for predicative adjectives and topic prominence in substrates of Singlish by Ansaldo (2009: 141–143). According to Bao, in order for the transferred feature to be productive in the emerging contact language, two further aspects are decisive: the frequency of usage in the input (Bao 2010: 796), and the so-called *lexifier filter*: the condition is that the “morphosyntactic exponence of the transferred system conforms to the (surface) structural requirements of the lexical-source language” (Bao 2010: 812). In order to identify the contexts required for successful transfer, detailed structural analysis of all languages involved is necessary (Bao 2010: 798, see also Siegel’s (2008) “functional transfer” and Poplack’s “structural conflict sites”, as discussed in section 2.2).

While the evaluation of typological distance or convergence between languages in contact is indispensable, it needs to be complemented by the socio-historical perspective. Thomason and Kaufman’s (1988) influential large scale study assesses the range of possible structural effects in various contact scenarios. They critically address postulations of limitations on transfer based on typological contrasts alone, and strongly link the extent of structural borrowing to extra-linguistic factors like amount of bilingualism and “cultural pressure” within a contact situation (Thomason and Kaufman 1988: 67). Their *borrowing scale* maps the intensity of contact (from casual to intense) between speakers to the extent of structural borrowing (Thomason and Kaufman 1988: 74–76).

Table 4.1: Thomason and Kaufman’s borrowing scale

Degree of contact	Effect on lexicon	Effect on structure
Casual	Content words (non-basic vocabulary)	None
Slightly more intense	Content words (non-basic vocabulary), function words	Minor structural borrowing: phonological features in loan- words only, extension of functions for existing syntactic structures
More intense	More function word, basic vocabulary, derivational affixes	Moderate structural borrowing: phonology, morphology, word order, syntax of clause-combining
Intense	Heavy lexical borrowing	“Anything goes, including struc- tural borrowing that results in major typological changes in the borrowing language” (Thomason 2001b: 71)

The extra-linguistic factors are both necessary, but by no means sufficient conditions to heavy structural borrowing, going as far as “contact-induced typological change” (Thomason 2001a: 1641).

Considering cultural pressure also means taking into account the different status of languages in contact situations. Uneven levels of power between speaker groups are reflected in the terms *substrate* and *superstrate* for the respective languages (Thomason 2001b: 75), this asymmetrical setting is common for English in historical contact situations (see also section 4.2.3).

Matras (2009: 312) stresses the need to distinguish between the *emergence* and *propagation* of structural innovations, identifying “directionality of bilingualism” and “norm-directed pressure on speakers” as main factors for the stabilisation of features. Although its exact role is contested, *code-switching* is widely acknowledged as a central mechanism of contact-induced language change (Thomason 2001b: 131–136). *Language shift*, i.e. the switch of a whole speech community to a new majority L1 after a prolonged period of bilingualism, commonly has far reaching structural consequences in the target language (Hickey 2010: 151). Additionally, *group SLA*, where learners, in the absence of a substantial native speaker community of the target language (TL), interact primarily among themselves, promotes “the preservation of contact-induced changes or innovations” (Winford 2003: 245).

The most radical result of multi-language contact in the absence of a single common language are mixed languages like *pidgins* and their nativised heirs, *creoles*, which have been described as “relexified” versions of substrate grammars (see e.g. Thomason 2001b: 159–160). The present work follows the assumption that, while languages are influenced by contact to widely varying degrees, there is more of a continuous scale than a radical division between contact varieties of a language and creoles of the same lexifier (see e.g. Lim and Ansaldo 2016, Mufwene 2001, 2008, Winford 2017). A case in point is Singlish, which has variously been described as “creoloid” (Platt 1975), or “almost a creole” (Gupta 1991), based on structural similarities with creole languages such as copula omission, or variable morphological marking of verb phrases, despite the absence of a preceding pidgin.

The importance of both *internal* and *external ecology* for language development is elaborated by Mufwene (2001). Using the analogy of *language evolution*, he describes the emergence of grammatical systems as a process of selection from a *pool of competing features*, contributed by the languages involved (Mufwene 2001: 4–6). The structural changes caused by external factors are thus strongly influenced by the internal ecology, or structural characteristics (Mufwene 2001: 192). Both aspects are linked inextricably in the speakers of the emerging language.

Acknowledging the central role of bilingual individuals in contact-induced language change entails the consideration of their *second language acquisition* (SLA) process. Yip and Matthews state that “[t]he developmental patterns in bilingual individuals parallel and reflect prominent features of contact varieties, such as Singapore Colloquial English, spoken by a community of adult bilingual speakers at the societal level” (Yip and Matthews 2007: 2). For individual SLA, external factors like “motivation, degree of access to and interaction with TL speakers, opportunities to use the TL, [and] attitudes to the culture” are decisive for the outcome (Winford 2003: 225). During early stages of SLA, the primary mode of communication is the “pragmatic mode”, exhibiting features of pragmatic languages (Matras 2009: 71, see also section 4.1.2).

Besides transfer from their L1, learners usually employ strategies associated with simplification such as “reduction of [target language] structures, rule regularization (via analogical levelling and (over-)generalization), and other strategies aimed at achieving ease of perception and production”, which are parallel to processes found in first language acquisition and internally motivated language change (Winford 2003: 217). Due to their lack of semantic contribution, the absence of inanimate and expletive pronouns in creoles is sometimes classified as a structural simplification (Siegel 2006: 21–22). A further parallel to first language acquisition is the influence of typological markedness on learnability:

typologically marked structures are acquired relatively late both by children and adult learners (Thomason 2001b: 52).

This section has sketched some general principles of grammatical transfer in different contact situations. It is obvious, as stated by Sharma (2012: 525), that “substrate, historical, or SLA forces are not entirely independent”, but rather exhibit intricate interactions. It has been shown that typological distance between languages is a possible, but not necessary obstacle for contact effects, as typological distance can in principle be overcome through intense contact, a high degree of individual and societal bilingualism, and suitable structural exponence in the lexifier. The optionality of overt subject pronouns in the substrate languages, and their status as a typologically marked feature, makes them a likely candidate for contact influence, especially in the case of expletive pronouns. The following section provides evidence how these processes manifest with regard to (structural conditions of) null subjects in second language acquisition, and in creole languages.

4.2.2 Null subjects revisited: Acquisition and contact

Contact with a non-NSL can lead to significantly increased amounts of overt subject pronouns in native speakers of null subject languages (e.g. Otheguy et al. 2010 on Spanish in New York, but see Silva-Corvalán 1994, who finds no such effect for Spanish in Los Angeles). This is analysed by Heine and Kuteva (2005: 47–48) as an extension from “minor to major use pattern” due to the influence of a diverging model in the contact language, in this case the higher frequency of overt subject pronouns in English as compared to the respective L1s.

In the same vein, studies on L2 Spanish (Liceras and Díaz 1999) and L2 Chinese (Li 2014) with speakers of various L1s (including English) show that learners with a non-null subject L1 background in principle observe the linguistic constraints of the target language the same way native speakers and speakers of other NSLs do, but significantly overuse overt pronouns, although less so with increasing proficiency. Structural constraints discussed in section 4.1.3 are confirmed for L2 Chinese: the two primary predictors for zero subjects are person and number, and reference continuity (Li 2014: 60). The prevalence of these factors is seemingly universal and robust. Other factors favouring zero are declarative clause, and nonspecific subject reference. As for L1 Chinese speakers, second person pronouns trigger a high amount of overt pronoun use for L2 speakers, third person inanimate contexts the lowest (Li 2014: 60). As expected, the presence of zero anaphora in the L1 leads to a higher rate of null pronouns in learner-Chinese of Japanese and Korean L1 speakers compared to native speakers of English and

Russian (Li 2014: 52, but see Chang and Zheng 2018 for conflicting findings for Japanese learners). This shows that in principle subject pronoun realisation is affected by transfer, and the opposite direction of influence, i.e. a higher rate of null pronouns in non-NSLs by NSL-L1 learners, can also be expected; given that overt subjects are the marked case, this effect can be expected to be even stronger.

This is shown in studies investigating the distribution of null pronouns in the L2 English of Chinese learners. The studies discussed here are generally based on experimental rather than naturalistic production data, not least to identify positive transfer, rather than the negative transfer evidenced by production errors (Tao and Healy 2005: 100). A recurring result is that speakers of radical NSL L1s fare better in grammaticality judgments of English sentences than speakers with canonical NSL L1s (Liu 2008: 276, see also Register 1990). Although expletive subjects are invariant in their absence in Chinese, raters do equally well for referential and non-referential subjects (Register 1990: 379). On the other hand, null subjects are ‘unlearned’ by speakers of radical NSL more readily than null objects, although there is no conclusive explanation for this asymmetry (see e.g. Chang and Zheng 2018, Hyams and Wexler 1993, Wang et al. 1992, Yuan 1997).

Both Yuan (1997) and Xiao (2002) are concerned with different subtypes of pro-drop in Chinese learners of English. In a set of grammaticality judgement tests, Yuan finds that learners acquire the ungrammaticality of null subjects in English much faster than for null objects, which is opposite to the development in English L1 acquisition (Yuan 1997: 487–488). In what can be interpreted as an instance of hypercorrection, Chinese learners reject null expletives more strongly than English native speakers, even in sentences like *Ø seems Mary is very tired*, which is rather common in colloquial spoken English (Yuan 1997: 482). On the other hand, when investigating production data, Xiao’s (2002) study on the development of subject-prominence in Chinese-English interlanguage of school-age children finds a much higher degree of null objects and null expletives (both 38%) compared to null referential subjects (10%; in contrast to the English control group with 5% null objects, 0% null expletives, 2.5% null referential subjects; Xiao 2002: 252). The usage of all three kinds of overt pronouns increases steadily over time (Xiao 2002: 262). Both studies conclude that topic prominence is a transferrable configuration in SLA, and can persist in learners’ language beyond an initial universal pragmatic mode of communication.

While SLA commonly involves a structured, guided learning environment and a strong norm orientation, this is not the case in creole genesis (Mesthrie and Bhatt 2008: 182). Null subjects in creole languages have enjoyed special attention from linguists, and their observations offer some testable hypotheses for the present study. The large scale investigation in the *Atlas of Pidgin and Creole Language Structures Online* (APiCS, Michaelis et al. 2013) allows for global statements

on pidgins and creoles, and provides evidence for the influence of NSL-substrates on their respective contact languages: the geographical division found in WALS (see section 4.1.1) is also mapped on the relevant APiCS feature (F62 “Expression of pronominal subjects”, Haspelmath and the APiCS Consortium 2013b); optional subject pronouns are thus largely restricted to contact languages in the Indo-Pacific and Asian region. Omission of referential subject pronouns is not a universal feature of contact languages: overall, less than one third of the APiCS languages (21/76) exhibits “optional pronouns” or “mixed behaviour”; among the English-lexifier pidgins and creoles included in the survey, only Chinese Pidgin English (4.24) and Singlish have optional subject pronouns.

(4.24) *Last week hap sellum only four thousand piece.*
 last week PFV sell only four thousand piece

‘The deliveries last week were 4,000 pieces.’

(Chinese Pidgin English, Li and Matthews 2013: Example 20-82)

Bislama shows “mixed behaviour”, allowing for null subjects only under certain conditions, i.e. restricted to one grammatical person (Meyerhoff 2000, see also section 4.2.4), a configuration alternatively termed “partial pro-drop” (Nicolis 2008: 279; see also section 2.1).

Another commonly observed split in the grammatical system of contact languages is that between referential and the more frequently attested non-referential null subjects.⁵ According to Nicolis (2008: 276), there is an “implicational scale of pro-drop” in contact languages, i.e. languages allowing for referential pro will also allow non-referential pro.⁶ APiCS provides two contexts for non-referential null pronouns, i.e. F63 “Expletive subject in *seem* constructions” (Michaelis and the APiCS Consortium 2013, with 23 languages out of 69 showing variable or null pronouns, 4.25), and the most common environment for zero pronouns in pidgins and creoles, F64 “Expletive subject of existential verb” (Haspelmath and the APiCS Consortium 2013a, 65 out of 75 sampled languages have variable or null expletives, 4.26).

⁵ Although some creoles also exhibit a “coexistence of null and overt expletives” (Nicolis 2008: 276), rated as “shared” in APiCS.

⁶ [+ref]pro > [-ref]pro correlation; [+ref]pro is a subcase of [-ref]pro; this is also shown in Gilligan’s 1987 cross-linguistic comparison.

- (4.25) *luk laik fiʃɪŋ ɡɔn bi gud ova hea.*
 look like fish-GER FUT be good over here

‘It looks like fishing will be good here.’

(Hawai‘i Creole, Velupillai 2013: Example 26-77)

- (4.26) *geʔ big kam faks*
 EXIST big kind shark-PL

‘There are big sharks (here).’

(Hawai‘i Creole, Velupillai 2013: Example 26-78)

Judging from the studies discussed above, native language affects pronoun realisation rates for learners, but in principle the same factors conditioning pronoun omission operate as in native speakers. This is slightly different in creole contact situations, where split systems of pro-drop based on person or referential status of the pronouns are encountered. Given the different classifications of the varieties investigated here, these differences can be expected to show in the present study as well.

4.2.3 English in contact: World Englishes

English as a global language has been adapted in its appearance in a wealth of diverse contact scenarios, and the principles of language contact as discussed in section 4.2.1 hold equally. The end of the British Empire has left a rich collection of local varieties with special political status, and by now, “second-language speakers of the language outnumber native speakers by far” (Schneider 2017: 35). It is no surprise that these new varieties of English have enjoyed considerable scholarly attention with regard to their historical genesis, current situation, and structural traces of contact. This section provides a short overview of the field of World Englishes studies, including conceptual frameworks, classifications of varieties and research approaches. More comprehensive accounts can be found e.g. in the monographs by Mesthrie and Bhatt (2008) and Schneider (2007), and the handbooks edited by Filppula et al. (2017) and Kachru et al. (2006).

Varieties of English affected by language contact are alternatively termed *New Englishes*, *Post-Colonial Englishes*, or *World Englishes*. The latter term is probably most neutral by now (Schneider 2017: 29), and will be used here. While much of the terminology in World Englishes studies is borrowed from creolistics, structurally and in terms of sociohistorical background there are a number of differences, although there is stronger overlap of language shift varieties and creoles

(Mesthrie and Bhatt 2008: 183, see also Mufwene 2001 for a critical discussion). Most World Englishes owe their formation at least partly to more institutionalised settings of language teaching, described as “multilingual scholastic English” by Gupta (1997). In contrast, Bao emphasises the continuous presence of the substrates, and thus an ongoing contact dynamic, as opposed to the “broken transmission” typically associated with pidgin and creole genesis, as the central distinctive feature of the emergence of World Englishes (Bao 2015: 4–6).

The most common classification of World Englishes is based on regional affiliation, i.e. English in Asia, Africa, etc. Another crucial classification parameter has been native language status, addressed e.g. by Quirk et al. (1972), distinguishing between English as a native language (ENL), English as a second language (ESL), and English as a foreign language (EFL). While essentially the same distinction is also represented in Kachru’s (1985) *Three Circles model*, its intent is more programmatic. Although as an abstraction this tripartite model necessarily misses the complexities of reality, such as internal linguistic heterogeneity of nations, its influence is undeniable. World Englishes study focuses on the Expanding and especially Outer circle, “norm-developing” varieties (Kachru 1985: 17). The *Electronic World Atlas of Varieties of English* (eWAVE, Kortmann and Lunkenheimer 2013) principally adopts Kachru’s tripartite classification scheme, with the additional distinction between “traditional” and “high-contact” L1 varieties (Trudgill 2002), and the inclusion of English-lexifier pidgins and creoles.

Schneider’s (2003, 2007) *dynamic model* is the most influential current theoretical model of World Englishes. He traces the historical development of various *Post-Colonial Englishes* and accounts for the linguistic results by common socio-historical processes in their formation. The selection and especially stabilisation of features from the feature pool, contributed by the languages in contact, is based on external factors like political status of a colony, and identity construction of local speakers by symbolic linguistic expressions (Schneider 2007: 28). Thomason and Kaufman’s 1988 borrowing scale (see Table 4.1 above) is reflected in the degree of structural nativisation in Schneider’s developmental phases of New Englishes (Table 4.2, adapted from Schneider 2007: 56).

Of these five phases, *nativisation* is the most dynamic, as it represents the phase where mutual accommodation of indigenous and settler population leads to novel means of expression. As described in section 4.2.1, the emerging variety develops additional structural possibilities, typically starting with a specific set of patterns. The interface of grammar and lexis is the most fertile ground for structural innovation, and high-frequency lexemes commonly act as the spearhead of linguistic change (Schneider 2007: 46). The final phase, *differentiation*, should not be confused with different degrees of proficiency as found in L2 environments, but rather refers to the formation of regional and social varieties, and

the development of a linguistic repertoire of fully competent native speakers (Schneider 2007: 53–54).

Table 4.2: Developmental phases in Schneider's evolutionary cycle of New Englishes

Phase	Sociolinguistics of contact	Structural effects
1: Foundation	minority bilingualism	incipient pidginisation (in trade colonies)
2: Exonormative stabilisation	spreading (elite) bilingualism	lexical borrowing
3: Nativisation	common bilingualism, toward language shift, L1 speakers of local English	heavy lexical borrowing, phonological innovations, structural nativisation: innovations at lexis-grammar interface, lexical productivity: code-mixing
4: Endonormative stabilisation	acceptance of local norm (as identity carrier), positive attitude, residual conservatism, literary creativity	stabilization of new variety, emphasis on homogeneity, codification
5: Differentiation	network construction (increasingly dense group-internal interactions)	dialect birth: group-specific (ethnic, regional, social) varieties emerge

Platt et al. (1984: 85–86) were the first to propose a list of shared morphosyntactic features across varieties, including the omission of morphological marking, marking of specific vs non-specific distinctions rather than definiteness, and unusual constituent sequences (for a more comprehensive account, see Mesthrie and Bhatt 2008, especially chapters 2 and 3). The large scale questionnaire based descriptions in Kortmann et al. (2004) and Kortmann and Lunkenheimer (2012, 2013) provide more empirical grounding for such generalisations. Kortmann and Szmrecsanyi (2004: 1189) identify irregular use of articles, the levelling of present perfect and simple past, a wider range of uses of the progressive, resumptive pronouns, the loosening of sequence of tense rules, invariant non-concord tags and lack of inversion in main clause questions as frequent features in contact varieties of English. However, the search for *Angloversals* (Mair 2003) has revealed that structural conformity can rather be found on more fine grained levels like *varioversals* or areal patterns of variation (Szmrecsanyi and Kortmann 2009, Kortmann and Schröter 2017).

Generally speaking, categorical differences between varieties of English are much rarer than quantitative differences in the usage rate of certain morphosyntactic patterns (Schneider 2007: 87). The importance of corpus linguistics for the investigation of such phenomena is obvious. As discussed by Mair (2017: 103), the advent of corpus linguistics, specifically the cooperative initiative of the ICE corpora, has considerably inspired World Englishes research. Using publicly available corpora has obvious advantages: it enables comparisons across varieties that are not feasible by an individual's data collection, and all results are transparent and testable. Often corpora are the only way to achieve enough instances of relatively low frequency phenomena or lexically determined variation. However, the price to pay is to give up control over the material sampled, including potentially relevant sociolinguistic information on speakers (see also section 1.3).

The central role of ecology in the genesis of World Englishes is evident. To investigate structural variation, large-scale approaches like eWAVE provide a good starting point to identify patterns worth further investigation; corpus-based structural investigations, following approaches like Sharma (2005), can further clarify the origins and present status of the features in question (see also section 4.4). The present study aims to contribute to a growing number of detailed corpus-based comparative studies trying to illuminate how the structural ramifications of contact present under varying sociolinguistic conditions.

4.2.4 Null subjects in (other) varieties of English

This section assesses the status of null subjects in varieties of English, i.e. their global distribution, and their occurrence in varieties other than those investigated here, which are discussed in dedicated sections below. The broad typological split between languages concerning the variable realisation of subject pronouns discussed in sections 2.1 and 4.1 is also a geographic one: “non-canonical” subject omission is widely attested in Asian languages and their contact varieties. A similar areal pattern can be detected for varieties of English.

eWAVE (Kortmann and Lunkenheimer 2013) shows that null subjects are found in a number of contact varieties of English. The questionnaire distinguishes between F43 referential (4.27) and F44 non-referential pronouns (4.28).

(4.27) Ø Haven't any rice (Sri Lankan English, eWAVE Example 752)

(4.28) Ø Must be getting late (Sri Lankan English, eWAVE Example 755)

F43 “Referential pronoun drop” is attested in 39 of all 76 varieties and rated “pervasive/obligatory” in eight of these. While the geographical division in eWAVE is not as obvious as in WALS and APiCS, pervasive use of referential null pronouns is especially prominent in Asian varieties (India, Singapore, Malaysia, Hong Kong) and creoles (Kriol, Torres Strait Creole, Gullah).⁷ Especially in contrast to the contact languages sampled in APiCS, expletive subject pronouns are surprisingly robust in varieties of English. F44 “Non-referential pronoun drop” is attested in 27 varieties overall, and pervasive in eight (again: Indian E, Singaporean E, Malaysian E, but this time *not* Hong Kong E; outside of Asia: Kriol, Bislama, Norfolk Island/Pitcairn E, Gullah, Hawai’i Creole; the last one is itself influenced considerably by varieties of Chinese). Both types of null subjects are uncommon in West-African Englishes (see Mesthrie 2012: 787), where substrates with optional pronouns are also rare (see Dryer 2013).

More detailed studies on null subject in varieties of English are scarce, but found e.g. in Mesthrie (1992) on South African Indian English (SAIE) and Meyerhoff (2000, 2009) on Bislama. Mesthrie investigates the distribution of null referential and pleonastic pronouns by lectal group in the language-shift variety SAIE. His comparison shows that deletion rates for referential pronouns decrease gradually with increasing stylistic level, from approximately 10% null subjects in basilectal, to 7.3% in mesolectal, and 4.6% in acrolectal speech style. A more fundamental shift is observed for zero dummy subjects: basilectal speakers omit more than half (57%) of all dummy pronouns, mesolectal about 20%, and there is no omission at all for acrolectal speakers (Mesthrie 1992: 170).

For Bislama, Meyerhoff (2000, 2009) states the absence of impersonal subject pronouns in expletives and raising constructions (see also Meyerhoff 2013). For referential pronouns, she identifies emerging inflectional marking via preverbal clitics as a licensing condition for null subjects in a split system of pronoun drop (underlined in 4.29).

- (4.29) *oli* *karem* *wan* *trak* *blong* *olgeta* *finis*
 AGR take INDF truck POSS 3PL COMPL

‘They’ve already got a truck for themselves.’

(Bislama, Meyerhoff 2013: Example 23-105)

⁷ Confirming observations made in section 4.2.2, object pronoun drop is less common in varieties of English (eWAVE feature F42, attested in 28 varieties).

The morphologically transparent third person verb forms permit null subjects more frequently than first and second person. Meyerhoff finds subject omission rates of 60% null for third person contexts, compared to 8% for first and second person combined (Meyerhoff 2000: 218).

The investigation of contact varieties of English is inconclusive on the status of overt expletive pronouns: while null expletives are attested across contact varieties, they are less common than referential null subjects. Furthermore, Mesthrie (1992) notices a categorical shift for acrolectal speakers of SAIE, interpreted as the result of successful parameter setting to the superstrate model in the acquisition process, and mirroring the hypercorrection observed for Chinese learners of English regarding the omission of non-referential subject pronouns (Yuan 1997, see section 4.2.2). The preference for third person zero in Bislama is explained by emerging morphological marking by Meyerhoff (2000), a grammaticalisation process that is not attested for any of the varieties discussed here. The remaining factors under investigation in the present study have not been systematically addressed for other varieties of English yet. The kind of contact, the amount and the depth of bilingualism are of crucial importance for the linguistic outcome (e.g. Thomason 2001a: 1640). The following sections therefore introduce the varieties under investigation not only with regard to their linguistic structure and ecology, but also with reference to their socio-historical background.

4.3 Asian varieties of English: Background and structure

While the three postcolonial varieties of English in India, Hong Kong and Singapore share certain characteristics, they also diverge in crucial aspects, both regarding their extra-linguistic and linguistic parameters. The following sections discussing the individual varieties are necessarily limited in scope. Their focus lies on the current role of English in the respective societies, the configuration of relevant substrate languages with regard to subject pronouns, and previous research on grammar, specifically null subjects, in the individual varieties.

4.3.1 English in India

The history of English in India starts around 1600 with the advent and rise of the East India Company. Following rapid economic and military expansion, the 18th century saw slowly increasing bilingualism in the local population (Bhatt 2004: 1017). British influence on the subcontinent culminated after the Great Rebellion in the establishment of direct governance of the Crown in 1858. Local

demand, as spelled out in Macauley's 1835 "Minute on Indian Education", led to the use of English in official and educational domains from the 19th century onwards, laying the seeds of nativisation (Sedlatschek 2009: 13–14). As in many colonial settings, knowledge of English became an indispensable asset for economic success and a marker of social status (Gargesh and Pingali 2017: 425). The importance of English remained contentious but unchanged by independence in 1947; its value as a symbol for educational and economic progress ultimately outweighed its association with the former colonial rule (Gargesh 2006: 90). Revoking plans of its abolishment, the permanent status of English as "associate official language of the Indian Union" was stipulated by the Official Languages Act in 1967 (Sedlatschek 2009: 19).

Currently English, alongside Hindi, is one of the two official national languages and almost the sole medium in the domains of higher education, science, technology, and all sorts of publications and media with a nationwide target audience (Gargesh 2004: 992). Furthermore, the position of English in multilingual India is secured by its function as an administrative link language of central government with non-Hindi states, as well as a *lingua franca* between speakers of different mother-tongue backgrounds (Gargesh 2006: 90).

Education policy reflects the societal and individual multilingualism and postulates the "three language formula", i.e. compulsory teaching and learning of the mother tongue + Hindi (or one other modern Indian language for Hindi speakers) + English (Sedlatschek 2009: 20), but in reality, the resistance against the respective regionally dominant local languages leaves English as the most widely accepted option. The linguistic model in language teaching is usually provided by locals, i.e. speakers of Indian English themselves, a crucial factor for nurturing nativisation (Schneider 2007: 166–167).

English is firmly an L2 in India (Mukherjee 2007: 182), the rather low number of self-reported L1 speakers has remained largely stable since independence (approximately 250,000, mainly from urban areas, and part of a well-educated elite, Sharma 2012a: 523). Speaker numbers are based on self-reports in census statistics. Obviously, English is a minority language in India (see Lange 2012: 54 for more detailed census data from 2001). Still, India is one of the largest English-using nations, current estimates of speaker numbers amount to approximately 100 million in a population of over 1 billion (Sharma 2012a: 523).

Pingali (2009: 6) describes English as "the language of the intellect and formality", which has a very limited role in private interactions of a more personal and emotional nature. Accordingly, Schneider (2007: 161–173) situates Indian English in between Phase 3 (nativisation) and 4 (endonormative stabilisation). Despite its flourishing use in creative and literary expression (see e.g. Paul 2003, Anjaria 2015), both Schneider (2007: 173) and Mukherjee (2007) doubt an immi-

nent switch towards a stronger role in identity construction which would encourage transition to Schneider's phase 5 (internal diversification).

It is misleading, however, to conceptualise the variety as a monolithic entity. *Indian English* as a cover term spans a wide range of second language varieties, "determined by L1s, region, socio-economic position, mode of acquisition, register of use, and attitude" (Sharma 2012a: 523). Shifts between different lectal levels and codeswitching are frequently encountered for highly proficient speakers (Sharma 2012a: 524). Bhatt (2004: 1017–1018) proposes a division between Vernacular and Standard Indian English, the latter with a strong orientation towards the British standard, while the former is more likely to show structural influence from local substrates.

Substrates of English in India mostly belong to four language families: Tibeto-Burman (a branch of Sino-Tibetan), Munda (a branch of Austro-Asiatic) and Indo-Aryan (a branch of Indo-European) in the North, Dravidian languages in the South (Gargesh and Pingali 2017: 430). The latter two families encompass the huge majority of speakers on the sub-continent (approximately 790 million, or 75% speakers of Indo-Aryan, more than 200 million, or 20% speakers of Dravidian languages, Lange 2012: 54, based on the 2001 census).

South Asia is one of the most linguistically diverse regions of the world and has a long history of multilingualism and intense language contact (Bhatia and Ritchie 2016: 156), resulting in the well-known *South Asian Sprachbund* (Thomason 2001b: 114–116). Morphosyntactic convergence within this linguistic area is attested extensively, including "special case marking of experiencers", SOV word order and the general absence of definite articles (Lange 2012: 56). The most relevant characteristic for the purpose of the present investigation is Moag and Poletto's account of "discourse level evidence", specifically their description of anaphora in spoken language (Moag and Poletto 1991: 235–241). Hindi and Malayalam serve as representatives of the two major languages families.

While the amount of morphological marking on verbs varies widely among the languages of the South Asian Sprachbund, zero anaphora is pervasive (Moag and Poletto 1991: 240, see also Krishnamurti 2003: 466–468 on Dravidian languages, Junghare 1988 on Indo-Aryan languages). Omitting subject pronouns is even described as "the unmarked condition in dyadic exchanges in South Asian languages" (Moag and Poletto 1991: 239), as illustrated below for Hindi and Malayalam (4.30, the same sentences are given in both languages). As shown in the description of Chinese in section 4.1, zero anaphora is permitted for familiar referents; in narratives, this can serve as a cohesive device, i.e. marking the absence of thematic boundaries (Moag and Poletto 1991: 237).

(4.30) **Question**

Hin. *tum ne use vah rumall di-yaa?*
 you he;OBL that handkerchief give-PST

Mal. *nii avanu aa kaileesu koTutto?*

‘Did you give him the handkerchief?’

Answer

Hin. \emptyset \emptyset \emptyset *de di-yaa*
 (I) (to him) (it) give-PST

Mal. \emptyset \emptyset \emptyset *koTuttu*

‘Yes, I gave it to him.’

(Moag and Poletto 1991: 236)

Non-referential pronouns are generally absent from Indian languages. The information-packaging function of e.g. existential constructions is fulfilled by flexible word order which allows for topicalisation of various syntactic arguments, including locatives (Lange 2012: 117–118, illustrated with 4.31 and 4.32 from Hindi). Widespread omission of referential subject pronouns, the absence of dummy subjects, together with verb-final word order, are among the most central criteria for topic prominence (Li and Thompson 1976: 467, see also section 4.1.2). The substrates of Indian English thus belong to the same topic-prominent type as Chinese; this is also assumed by Lange (2012: 59).

(4.31) *pustak mez par hai*
 book table on is

‘The book is on the table.’

(Lange 2012: 118)

(4.32) *mez par pustak hai*
 table on book is

‘There is a book on the table.’

(Lange 2012: 118)

It is no exaggeration to say that scholarly investigation of Indian English initiated the World Englishes enterprise, most notably publications by Kachru (especially Kachru 1965, 1982, 1986). Following early accounts in the form of lists of

errors and deviances from the British standard, the question of the legitimacy of “Indian English” has been of great interest (e.g. Gargesh 2006, Schneider 2007: 171–172). Recent years, and the advent of corpus linguistics have given way to broader investigation of nativisation on all structural levels and allowed for more system-internal descriptions (e.g. Lange 2012, Pingali 2009, Sedlatschek 2009, Sharma 2012a).

Unlike accent, where regional dialects like *Marathi English*, *Hindustani English* or *Kannada English* are encountered (Gargesh 2004), Bhatt (2004: 1016) states that “Standard Indian English is essentially similar in its core syntax to Standard British English”. This is shown by Balasubramanian’s (2009) analysis of register variation. Due to the extensive morphosyntactic convergence of the substrates, differences between regional or L1 varieties of Indian English are mainly accounted in phonology and pragmatics (Hickey 2010: 525).

Concerning its morphosyntax, Indian English is firmly an L2 variety (Lunkenhoimer 2012). Schneider (2007: 169) notes that many morphosyntactic features of Indian English are shared with other World Englishes, such as “invariant tags, pluralisation of mass nouns, omission or insertion of articles, use of progressive with stative verbs, *wh*-interrogative clauses without inversion, reduplication of adjectives and verbs, a wider range of uses of the past perfect structure”. The most innovative structures are found in lexico-grammar, especially verb complementation patterns, as evidenced by various corpus-based studies, e.g. Mukherjee and Hoffmann (2006), Mukherjee and Gries (2009) and Schilk (2011).

The complexity of substrate influence in Indian English is analysed by Sharma for article use (Sharma 2005) and the aspectual system (Sharma 2009). Her meticulous investigation relativises the amount of transfer, and alerts to the interplay of substrate features and universal discourse principles. One of the crucial factors for article omission, discourse familiarity, is also relevant for the present study. Article realisation is guided by the economical principle of disambiguation, leading to higher omission rates for NPs denoting unambiguously identifiable referents (Sharma 2005: 558).

Lange (2012) provides a detailed account of syntactic phenomena of educated spoken Indian English as represented in ICE-India. The focus lies on the “discourse-pragmatic sentence structure” (Lange 2012: 90), i.e. the syntactic representation of topicalisation and focus, dislocation, and cleft constructions. Structural features distinctive for Indian English include non-initial existential *there* constructions (see 4.35) and *only* as presentational focus marker, while higher frequencies for topicalisation and left dislocation constructions “are supposedly common to the New Englishes” (Lange 2012: 237, see also Winkle 2015: 106). Overall, rather than vaguely marking “some diversion from the canonical topic-comment structure”, the common motivation for the constructions in ques-

tion is to create textual cohesion, a function that can also be fulfilled by zero anaphora (see section 4.1.2). Lange interprets this pragmatic shift as a contact effect from the “pragmatic Sprachbund” substrates of Indian English (Lange 2012: 238–240)

Among the morphosyntactic features of Indian English deemed “likely to arise out of substrate transfer” by Sharma are the omission of referential (4.33), non-referential (4.34), and object pronouns (4.33; see also eWAVE features F42–47, Sharma 2012a: 524). Bhatt (2004: 1027–1029) rates both subject and object drop as systematic features of vernacular Indian English, especially for “topic-connected” arguments.

- (4.33) A You got tickets?
 B No, Ø sold Ø already.
 (Bhatt 2004: 1023)

- (4.34) During monsoon we get lot of rain and then Ø gets very soggy and sultry.
 Ø rained yesterday only.
 Here Ø is not safe to wait
 (Bhatt 2004: 1028)

However, the existential construction presents as a form specific to Indian English: dummy *there* is substituted by post-verbal adverbial *there* (4.35, Bhatt 2004: 1029)

- (4.35) Bread is there
 ‘There is bread.’
 (Trudgill and Hannah 1985: 109)

Lange’s corpus analysis shows that this phenomenon is robust in educated spoken Indian English, but its function is rather different from the canonical existential: it is hardly used for introducing new referents, but rather used as a cohesive device, i.e. as a “topic repetition strategy” (Lange 2012: 106–107). Concerning the prevalence of null subjects in Indian English, Bhatt remarks on a strong stylistic division:

[...] whether a subject will be dropped or not depends on, among other things, the formality of the context: in less formal contexts, the probability of subject-drop is high, close to 100 percent, whereas in formal, and especially in the written mode, the probability of subject-drop is very low, close to zero. (Bhatt 2004: 1017)

It will be interesting to see how this phenomenon behaves in the sample of spoken educated Indian English analysed in section 5.3.

4.3.2 English in Hong Kong

English took settlement in Hong Kong in the 19th century, when Britain obtained direct rule via the treaty of Nanking in 1842 (Setter et al. 2010: 4). Before that, communication between traders and locals had mainly been conducted in various forms of Chinese Pidgin English, now extinct (Setter et al. 2010: 104, Bolton 2003: 178–189). Missions' endeavours, most prominently the opening of mission schools, initiated the spread of English (Bolton 2003: 192–194). The presence of English was strengthened and stabilized by the 99-year lease granted to Britain in 1898, and its establishment as the sole language of government (Schneider 2007: 135). This also enabled steady trade activity, increasing the permanent population, and the rise of Hong Kong as a focal point for banking and commerce (Setter et al. 2010: 4). Especially the late colonial period from the 1970s onwards saw rising wealth among larger parts of the population, establishing compulsory education and introducing bilingualism on a scale beyond a small elite via the education system (Schneider 2007: 136–137). The role of English as a gatekeeper to professional advancement evoked widespread demand by parents for its use as medium of instruction from secondary level onwards since the 1980s, leading to a rise in English knowing bilingualism by the end of the century (Evans 2016: 303–304). For the time of data collection for ICE-Hong Kong in the late 1990s and early 2000s, Bolton (2003: 87) reports an amount of 43% of the population knowing English.

Unlike other post-colonial varieties of English, Hong Kong English did not gain momentum as a side effect of political independence, as the former Crown colony merely switched rulers to the PRC in the handover 1997 (Schneider 2007: 136). Granted the status of a *Special Administrative Region* (SAR) under the “one country, two systems” framework for 50 years after the handover, a local identity, rooted in, but distinct from mainland Chinese culture, has emerged (Hyland 1997: 207). However, how the role of English develops in the future, especially after the end of this transition phase, remains to be seen.

Throughout its history, the population of Hong Kong was and is rather homogeneous ethnically and linguistically: approximately 96% of its 7 million inhabitants (as of 2008, Setter et al. 2010: 2) is ethnically Chinese and the vast majority indicates Cantonese, a Southern Chinese Yue dialect, as their “usual language” (Wong 2017: 10). Unlike in India or Singapore, this monolingual setting largely eliminates the need for English as a means of intranational communication (Setter et al. 2010: 5).

Although Hong Kong continues to recognise English as a secondary official language (alongside the “conveniently ambiguous [term] *Chinese*”, Evans 2013: 306), its use is now largely confined to written registers and restricted to second-

ary and higher education, higher courts, and international companies (Yip and Matthews 2007: 10). Together with Mandarin Chinese (also termed *pǔtōnghuà*, ‘common speech’), English is perceived as a tool for international communication but evokes no emotional attachment (Gisborne 2009: 153). Education policy remained autonomous since the handover and aims at full functional biliteracy (in Chinese and English) and trilingualism (in Mandarin, Cantonese and English). Given the restricted domains of use, especially for spoken English, this goal is probably not realistic for the population at large (Evans 2013: 307). Indeed, self-reported native speakers of English only form a tiny minority, and there is no significant shift in this direction either. As “additional languages”, both Mandarin and English are on the rise, mostly at the expense of “other Chinese dialects” such as Hakka and Southern Min (Evans 2016: chapter 5).

Still, Hong Kong English is now widely accepted to have risen beyond the ephemeral “English in Hong Kong” of earlier decades (Wong 2017: chapter 1), as evidenced especially by Bolton (2002), including a discussion of distinctive features and creative uses of the variety, both common indicators of an emerging variety; terming it a “stillborn variety” (Pang 2003) was certainly premature. Schneider (2007: 133–139) accordingly locates Hong Kong English in phase 3 of his dynamic model, presenting the increasing “complaint tradition” as language external, and the emergence of structural nativisation in morphosyntax as truly linguistic evidence.

Southern Chinese dialects of the Yue family, Cantonese specifically, form the dominant substrate of English in Hong Kong. Cantonese and related dialects are spoken languages primarily. Although the handover in 1997 strengthened the role of Mandarin in Hong Kong and proficiency is increasing, primary education is still conducted in the local, rather than the national language (Evans 2016: chapter 5).

Differences between Cantonese and Mandarin phonology are strong enough to render them mutually unintelligible; in terms of morphosyntax, however, structural similarities prevail (Matthews and Yip 1994: 5). Both are topic-prominent (see e.g. Yip and Matthews 2007: 135), and with regard to the features in question here, as elaborated in section 4.1.1, they are similar enough to allow for discussion under the cover term “Chinese”. More divergent structurally are dialects from the Min branch such as Hokkien and Chaozhou (alternatively transliterated Chiu Chau). However, while they are present especially among older speakers in Hong Kong, their influence is rather limited due to their lack of prestige and status as minority dialects with further diminishing speaker numbers, described as lack of “ecological nurturing” by Yip and Matthews (2007: 11; concerning the role of Min dialects in Singapore, see section 4.3.3).

A typologically remarkable feature of Chinese are pre-nominal relative clauses, which are rare in general, and even more exceptional in SVO languages

(4.36, Yip and Matthews 2007: 156). Yip and Matthews remark on a specific pattern of relative clauses in informal spoken Cantonese, as the most common input for L1 acquisition, the “classifier relative” (4.37).

(4.36)	Man.	<i>Wǒ</i>	<i>xǐhuan</i>	<i>de</i>	<i>yīfú</i>	<i>hěn</i>	<i>guì</i>
	Can.	<i>Ngó</i>	<i>zūngji</i>	<i>ge</i>	<i>sāam</i>	<i>hóu</i>	<i>gwai</i>
		I	like	PRT	clothes	very	expensive

‘The clothes I like are expensive.’

(Yip and Matthews 2007: 159)

(4.37)	<i>Ngó</i>	<i>zūngji</i>	<i>gó</i>	<i>dī</i>	<i>sāam</i>	<i>hóu</i>	<i>gwai</i>
	I	like	DEM	CL	clothes	very	expensive

‘The clothes I like are expensive.’

(Spoken Cantonese, Yip and Matthews 2007: 159)

Yip and Mathews provide this structure, which is deviant from both Mandarin and other Southern Chinese dialects, as a possible source of null relative pronouns in Hong Kong English.

The seminal study on Hong Kong English by Luke and Richards (1982) is still firmly grounded in the colonial setting and describes a learner or performance variety that was acquired anew by each generation of speakers, an exonymatively oriented “English in Hong Kong” rather than the distinct variety “Hong Kong English”. The dynamic situation around the end of colonial rule and the handover has inspired a number of studies on language attitudes and policy, e.g. Bolton (2011), and Evans (2009, 2016). The chapters in Bolton (2002) are concerned with various aspects of Hong Kong English, including its creative uses, providing evidence for an emerging variety. Yip and Matthews (2007) discuss issues of first and second language acquisition, and how they manifest structurally in children growing up bilingual in Hong Kong. They detect parallels of early stages of “unbalanced bilingualism” with the development of contact varieties like Singapore English (Yip and Matthews 2007: 2).

Wong (2012) provides an overview of morphosyntactic features of Hong Kong English. The grammar is characterised by simplification and deletion phenomena, like lack of morphological marking for NP (plural) and VP (third person, past), levelling processes like the regularisation of irregular past forms, plural marking on non-count nouns, and irregular use or absence of gender distinction in pronouns. Most of these features can be traced back to the analytic substrate

structure; on the other hand, many such simplification and regularisation processes are typical for L2 varieties in general (Lunkenheimer 2012).

While most studies of Hong Kong English focus on sociological, attitudinal and educational issues, structural features attracting specific attention include relative clauses (Gisborne 2000, 2009) and discourse phenomena like cleft constructions (Yao 2016), and expressions of gratitude and code-mixing (Wong 2017). Gisborne (2009) presents a systemic approach to Hong Kong English grammar, identifying “lack of finiteness” as a larger typological setup transferred from the substrate. This larger parameter possibly accounts for copula omission and difficulties in distinguishing lexical categories, especially the contrast between verbs and adjectives, which can act as predicates (Gisborne 2009: 155).

In their structural account of Hong Kong English, Setter et al. (2010: 56–59) identify several features of topic prominence, such as the double NP construction, and a wider range of possible topics compared to Standard English (4.38, see also section 4.1.2). Unlike in Standard English, where the topic NP is usually marked periphrastically by prepositions like *regarding* TOPIC or *as for* TOPIC, Hong Kong English frequently mirrors the topic-prominent structure of Cantonese both in speech (4.38) and writing (4.39).

(4.38) Vancouver they have high-rise buildings they have... relatively good food
(Spoken HK-English, Setter et al. 2010: 57)

(4.39) Passengers who take the ferry service from Ma Liu Shui, they can enjoy
a free ride from Tap Mun to Wong Shek.
(Written HK-English, Setter et al. 2010: 57)

But Setter et al. (2010: 58) state that while Hong Kong English shows these properties of topic-prominent languages, sentences without overt subject are rare in their spoken data set. Notable exceptions are relative clauses, where subject pronouns are omitted more freely (4.40, eWAVE feature F193). Null relative pronouns are a frequent feature in Chinese learners of English as well (Hung 2012).

(4.40) This is the student Ø did it
(Gisborne 2000: 359)

This phenomenon is so frequent in Hong Kong English that it is commonly found even in formal writing (Gisborne 2000: 359), but less common in other Asian Englishes. Besides this preference for zero-subject relative clauses, a further remarkable difference is the attested absence of null dummy subjects in Hong Kong English in the eWAVE ratings. While this seems to negate the influence of referential status on pronoun realisation, this assumption is countered by feature F46,

“deletion of *it* in referential *it is*-constructions”, and feature F47, “[...] in non-referential *it is*-constructions”, respectively: while no information is available for referential *it*, deletion of non-referential *it* in this specific context is pervasive for Hong Kong English (Wong 2013).

4.3.3 English in Singapore

The first accounts of Singapore as a settlement precede the arrival of English, which can be dated to 1819, when Sir Stamford Raffles obtained the rights to establish a free port in the strategically fortunate location at the southern tip of peninsular Malaysia. Less than 10 years later, the former fishing village had developed into a thriving commercial hub with more than 10,000 inhabitants, mainly (Straits) Chinese, who have formed the majority of the population ever since, but increasingly attracting traders and labourers of diverse backgrounds. As part of the Straits settlements, Singapore acquired the status of Crown colony in 1867 and continued to prosper, accommodating not only a small European (less than 10% of the population), but also a growing Asian elite of Chinese, Malay and Indian professionals (Schneider 2007: 153–154). The early 20th century saw a large influx of Chinese children to English-medium schools, a likely starting point for nativisation (Gupta 1998: 114–116). The colonial tradition was broken by Japanese occupation during WWII (1942–45), and after a short-lived membership in the Federation of Malaya, independence was declared in 1965 (Schneider 2007: 155).

The following decades witnessed stunning economic growth and successful nation building, steered by the perennially governing People’s Action Party (Schneider 2007: 155). Their hands-on style of government includes active language policy, designed to further encourage a distinct Singaporean identity built on the local cultural and ethnic mix, as well as global economic competitiveness. Official languages are chosen to represent the major ethnic groups, i.e. Tamil for Indians (approximately 8% of the population), Malay for Malays (approximately 15%, additionally designated the “national language”), Mandarin Chinese for Chinese of various backgrounds (approximately 75%), and English as the ethnically neutral lingua franca and “working language” (Lim and Foley 2004: 5).

This twofold policy of strengthening the local roots as a basis for Western orientation is also reflected in the bilingual education policy. Each child receives education in English as a mother tongue and in their respective mother tongue. The latter is assigned by the ethnicity of the parent(s), regardless of the actually spoken languages, weakening the usefulness of the indigenous languages and strengthening the position of English instead (Schneider 2007: 157–157). Since

1979, the Speak Mandarin Campaign has steeply increased proficiency in Putonghua at the expense of traditionally indigenous Southern and contact varieties of Chinese (“Chinese dialects” as home languages slide from 79% in 1980 to 19% in 2010, Bao 2015: 33, for recent data on the rising use of Mandarin Chinese, see also Siemund et al. 2014). Currently, English is firmly established as the primary language of modern Singapore, and increasingly used as the “most frequently spoken home language” by about one third of the population (2010: 33% of Chinese, 17% of Malay, and 42% of Indians; for more detailed census data, see Bao 2015: 33).

To the displeasure of its government, Singapore English has a distinct local flavour; consequently it has been the target of another political initiative, the *Speak Good English Movement*, pointing out perceived common mistakes in accent, vocabulary, orthography and grammar via a yearly campaign and its own newspaper column (Rubdy 2001). However, Singaporeans value their local language, illustrated most notably by the inception of the *Coxford Singlish Dictionary* and the *Speak Good Singlish Movement* (Wee 2014).

While early accounts of Singapore English assumed a lectal continuum, akin to a post-creole continuum (e.g. Platt 1975, Pakir 1991), Gupta (1994: 7–9) analyses the stylistic repertoire as a diglossic split between standard and vernacular Singapore English. Current speakers, especially among higher socioeconomic strata, as well as the younger generations, can be assumed to have native-like competence, including command over a range of registers with more or less localised features (for an analysis of this variation in the *cultural orientation model*, see Alsagoff 2010, Leimgruber 2009, 2013). Consequently, Singapore English is considered an L1 in eWAVE, and firmly in Schneider’s phase 4, with hints of a continuing development towards phase 5 (internal diversification).

From its very beginnings, Singapore has been a multi-ethnic and multilingual society. At present, the dominant languages are the Sinitic language Mandarin Chinese, the Austronesian language Malay, or Bahasa Melaya, very similar to Malaysian (Percillier 2016: 2), and the Dravidian language Tamil (see also section 4.3.1). However, besides the current language situation, it is important to take into account the linguistic ecology at earlier stages of variety formation. Lim (2009) provides a detailed overview of historical substrates and their influence on the most famous feature of Singlish, the discourse particles.

Substantive influence of the Indian languages on Singapore English is unlikely, given the small share of ethnic Indians in the population. Historically, the substrate in Singapore was formed by Chinese dialects from the Southern

Min branch, most prominently Hokkien and Teochew⁸, but also Cantonese from the Yue branch, dialects from the Hakka branch, and Hainan Min, rather than Mandarin (Bao 2015: 18). Especially Hokkien was widespread in Singapore until political support for standard spoken Chinese Putonghua pushed it out of the linguistic repertoire of younger speakers (Rubdy 2001: 341).

Malay was represented mainly by contact varieties like Baba Malay, the Malay-based creole of the Peranakans, and Bazaar Malay, the pidginized lingua franca of Malaya, both of them considerably influenced by Chinese varieties themselves (see e.g. Ansaldo 2009: 21; see Aye 2006 on Bazaar Malay, Lee 2014 on Baba Malay). Both Bazaar Malay and Baba Malay are topic-prominent (4.41, see also Lim 2011: 283).

(4.41) *Itu barang ah*
that goods PART

ada kali mahal, ada kali murah itu pasar sana
have time expensive have time cheap that market there

‘As for those goods, sometimes expensive, sometimes cheap, that market.’
(Bao and Aye 2012: 162)

According to Bao (2015: 30), the founder effect of Baba Malay on early Singapore English is mitigated by the continued presence of the shared substrates (but see Sato 2013, citing partial *wh*-movement as evidence for lasting Malay influence on Singapore English). In addition, the Chinese population does not only form the large majority, but also shows a stronger tendency for language shift than the Malay population (Siemund et al. 2014: 343). Bao (2015: 35) states that “[i]ntense scholarly research during the past fifty years or so has demonstrated clear and irrefutable Chinese influence on the grammar of Singapore English”. In the end, all substrates converge with regard to subject omission in general, but diverging influence is expected in the emergence of variety specific contexts for null subjects.

Despite its minor size, the densely populated island nation of Singapore harbours an exceptionally dynamic linguistic environment and its nativised variety, affectionately termed Singlish, has consequently inspired intense academic attention as a case study of urban multilingualism and an illustration of global developments in a very condensed environment. Hence, research on

⁸ The alternative transliterations Chaozhou and Chiu Chau are used in Hong Kong, see section 4.3.2.

English in Singapore has a comparatively long history among the New Englishes. Early approaches are mainly concerned with lectal variation and the development of a local standard, e.g. Crewe (1977), Platt (1975, 1979), Richards (1977), and Tongue (1979) as the initial descriptions in the 1970s, followed by numerous edited volumes and monographs in the 1980s and 1990s discussing more diverse aspects, such as the role of different languages in society and education (e.g. Foley 1988, Foley et al. 1998, Gupta 1994), and several detailed structural investigations by Platt and Weber (1980), Platt et al. (1983), and especially Ho and Platt (1993), who apply the variationist method to the detailed analysis of copula deletion in Singapore English. More current publications with a focus on Singapore English grammar include Deterding et al. (2003), Deterding (2007), Leimgruber (2013), Lim (2004), and most recently Bao (2015).

Substrate influence in Singapore English is attested on all linguistic levels, and encountered in different speech styles as well (see e.g. Leimgruber 2009, Schröter 2010). Differences between ethnic groups are mainly found in phonology, especially intonation (Lim 1996). Feature-based comparison across varieties of English based on the eWAVE feature set shows that Singlish exhibits numerous rare or even unique features that are not found in (m)any other varieties, such as *ever* as experiential perfect marker (F108), and the *give*-passive (F153), but also features typical of World Englishes such as BE deletion, absence of verbal marking, and article omission (Schröter 2012). Complementary to acceptability ratings, Bao (2015: 131) distinguishes between *productive* and *unproductive* transfer features, based on their actual usage frequency. On the most colloquial level, substrate influence is pervasive, leading Ansaldo (2004: 139) to claim that “though considered a variety of English, Colloquial Singapore English is really typologically closer to languages of East and Southeast Asia.” Compared to other World Englishes, there is an extensive body of research on specific structural features of Singlish and their basis in local substrates, e.g. discourse particles (Lim 2007, Platt and Ho 1989, Wong 2014: chapter 7), the aspectual system (Bao 2005), and passive constructions (Bao and Wee 1999, Kim and Sato 2013).

The investigation of BE deletion by Ho and Platt (1993) is one of the earliest systematic studies of grammatical variation in World Englishes via multivariate analysis. BE deletion in different grammatical contexts is widespread in Singlish, and is shown to be indicative of educational level and proficiency in English (Ho and Platt 1993: 54–56). In contemporary Singapore English, BE deletion still acts as a sociolinguistic marker, evidenced by the high degree of stylistic variation (Schröter 2010).

Besides the aspectual system, grammatical replication, as described by Heine and Kuteva (2005), is found for the Malay-based *kena*-passive, relative clauses with *one*, question formation with *or not*, and existential *got*. The latter

two phenomena are also relevant for null subjects in Singlish. The *kena*-passive in Singlish (4.42) is an exact calque of the Malay construction (4.43). It is based on the lexeme *kena* ('to strike, to come into contact with'), which is compatible with both positive and negative outcomes. In its function as a passive marker, however, it carries an adversative connotation, which is reflected in its usage in Singlish as well (Bao and Wee 1999: 4).

(4.42) I *kena* sexual harassed again you know

(Bao 2015: 182)

(4.43) *Budak jahat itu kena pukul*
boy naughty the kena beat

'The naughty boy was beaten.'

(Bao and Wee 1999: 4)

Although Singlish does not share the strong tendency for null relatives reported for Hong Kong English (section 4.3.2, but see 4.48 below), it does exhibit a specific structure constituting a replicated substrate construction. These relative clauses provide an alternative to the canonical English pattern. They occur with or without overt English relative pronouns and use clause-final relative particle *one*, most likely modelled on Mandarin Chinese *de*, but preserving the post-nominal ordering of English (on other functions of *one* as nominaliser, pronominal and emphasis marker, see e.g. Bao 2009, Wee and Ansaldo 2004). Zero relativiser without *one*, however, is deemed ungrammatical by Alsagoff and Ho (1998, 4.44). Following the Chinese pattern, Singlish *one* relative clauses can even occur without heads (4.45).

(4.44)	The man who sell ice-kachang <i>one</i>	gone home already
	The man who sell ice-kachang	gone home already
	The man Ø sell ice-kachang <i>one</i>	gone home already
	*The man Ø sell ice-kachang	gone home already

(Alsagoff and Ho 1998: 133)

(4.45) Ø Don't have car *one*, I don't want.
'I don't want [a man] who does not own a car.'
(Alsagoff and Ho 1998: 135)

The *X or not* construction is a structure frequently encountered in Singapore English to express *yes-no* questions. It is commonly formed without overt subject pronoun (4.46). The structure can be traced back to the Min dialects Teochew and

Hokkien (Matthews 2010: 764, 4.47), its occurrence in Mandarin and Cantonese is much more limited (Yip and Matthews 2007: 11).

- (4.46) Can Ø answer the question or not?
 ‘Do you know the answer to the question?’
 (Wee 2004: 1064)
- (4.47) *U tsi a bo?*
 have money or not
 ‘Do you have money?’
 (Teochew, Yip and Matthews 2007: 11)

Singlish also possesses an alternative existential construction fulfilling the same presentational function as standard English existential *there* BE (4.48, see also Lee et al. 2009). The Singlish existential *got*-construction shows obvious structural parallels to the Chinese existential *yǒu*-construction: it appears without an expletive subject pronoun and is immediately followed by the NP establishing the new topic (line 1; see also section 4.2.1 above). Subsequent reference is possible by zero (line 2) or overt pronoun (line 3). *Got* is also used in a stative possessive sense in Singlish, usually in connection with an overt subject (4.49).

- (4.48) 1 Then that time *lah* got [two teachers]_i
 2 Ø_i start loving each other
 3 Then they_i marry one year later <ICE-SG:085#448–449:A>
- (4.49) B Anybody got calculator
 A I think I have I have a calculator <ICE-SG:017#106–107>

Null pronouns in subject (4.50) and object position (4.51) are already mentioned in early accounts of Singapore English, e.g. Platt et al. (1984). The condition for null subjects is usually described as “familiarity” of the referent. Gupta states that “where the subject can be retrieved from the context, Colloquial Singapore English does not require it to be expressed” (Gupta 1994: 10, see also Leong Ping 2003).

- (4.50) Dis Australians_i, you see dem hold hand hold hand,
 honey here, honey there, darling here, darling dere,
 next moment Ø_i separated already.
 (Platt et al. 1984: 155)

- (4.51) In Australia, people never carry umbrella_i –
 so if you carry Ø_i they will laugh at you.
 (Platt et al. 1984: 77)

Instances like these are commonly mentioned as central proof for extensive substrate influence and even taken as an indicator of a *typological split* of Singlish from its lexifier, based on the distinction between subject-prominent and topic-prominent languages (e.g. Ansaldo 2004, 2009, Tan 2003). Bao and Lye (2005) investigate the *bare conditional* (4.52) as evidence for the transfer of topic-prominent features beside null subjects.

- (4.52) (sisters talking about a dish)
 Don't care *lah*. Ø Want to eat, eat; Ø don't want to eat, then don't eat.
 'Don't worry.
 If you want to eat it, eat it; if you don't want to eat it, then don't eat it.'
 (Bao and Lye 2005: 283)

Bao and Lye (2005: 287) show that “[i]t is clear [...] that the entire cluster of properties associated with topic prominence is transferred from Chinese to Singapore English”, including bare conditionals and Chinese-style topics. Null subjects are also described for lexically specific constructions, e.g. Wee (2003a) on *know* (4.53), and Wee (2003b) on *don't know* as emerging discourse markers, and in combination with modal verbs, as described by Bao (2015: chapter 7) for *must* and Wong (2014: chapter 5) for *can/cannot*.

- (4.53) The bus is late Ø know
 (Wee 2003a: 7)

Can/cannot are ubiquitous Singlish expressions (Wong 2014: 139). *Can* is employed as a single-word response (4.54), or as question tag (4.55), sometimes in combination with *or not* (4.56). Due to their functions as invariant particles or tags, such instances with *know* and *can* are not included in the quantitative analysis.

- (4.54) A Let's have lunch on Thursday or Friday
 B Ø Can.
 (Wong 2014: 161)
 (4.55) Meet me tomorrow, can?
 (Wong 2014: 161)

- (4.56) I want to go home, can or not?
 ‘Can I go home?’
 (Wee 2004: 1064)

Null pronouns are a salient feature of Singapore English, and Schröter (2010) shows that subject omission is a stable feature of contemporary educated Singapore English, even across speech styles (see also Deterding 2007: 58). Commonly observed contexts for null subjects in Singapore English are *X or not* questions, *got*-existentials, the bare conditional, and modal verbs, especially *can* and *must*. Based on previous analyses, the favouring influence of reference continuity on null subjects is likely. However, in spite of the comparatively rich research on Singlish null subjects, Bao’s (2015: 190–191) demand for systematic quantitative investigation of the phenomenon is unfulfilled thus far. The analysis conducted in chapter 5 incorporates these observations to test their influence on the realisation of subject pronouns in Singlish. Further methodological preliminaries for this comparative analysis are discussed in the following section.

4.4 Comparing varieties

While the evidence for *Angloversals* remains inconclusive so far, the detection of areal patterns and especially *varioversals* in the morphosyntax of varieties of English has been more successful. The following section 4.4.1 provides arguments for analysing Asian Englishes as a subgroup of global varieties of English, followed by a methodological primer (section 4.4.2) for the comparative variationist study in chapter 5.

4.4.1 Asian Englishes

Contact varieties of English in South and Southeast Asia, the *Asian Englishes*, have been a prosperous field of study for years now, see e.g. the establishment of the eponymous journal in 1998, the book series *Asian Englishes Today* including e.g. Kachru (2005) and Kachru and Nelson (2006), and the edited volumes by Bolton and Kachru (2007), Kirkpatrick and Sussex (2012), Lim and Gisborne (2011), Low and Hashim (2012), Murata and Jenkins (2009), Prescott et al. (2007) and Wee et al. (2013), often integrating political perspectives and a focus on cultural and socio-historical background with linguistic research. This and other similar types of subgrouping based on region or presumably shared socio-historical background have also been discussed critically (e.g. by Ansaldo 2010, Lim

2009). Structurally, however, the relevant substrates converge with regard to the variable expression of subject pronouns, as shown in section 4.1.1, but also in the description of the more specific linguistic background of the individual varieties in the relevant subsections of 4.3.

As exhibited by parallels in the structural descriptions of the three varieties above, the status of Asian Englishes as a legitimate unit of analysis is also justified from a synchronic point of view. This is confirmed by large scale comparative investigations, such as Kortmann and Lunkenheimer (2013), which additionally include descriptions of the Asian varieties Malaysian, Pakistani, Sri Lankan and Butler English.⁹ In their global synopsis, Kortmann and Wolk (2012: 921–923) show with the help of *NeighborNet* visualisation that, except for the pidgin Butler English, all Asian Englishes are positioned within a group of L2 Englishes, together with African L2 varieties (i.e. Platt et al. 1984's New Englishes). Stronger ties are detected between the Southeast Asian varieties and Indian English, while the remaining South Asian varieties are more diverse structurally (Kortmann and Wolk 2012: 921). The network shows that, on the one hand, variety type is the most decisive predictor of the structural profile of varieties based on the WAVE feature catalogue. This is also visible from the clusters of L1 varieties, and pidgins and creoles. On the other hand, the network proves the unifying effect of areality. This is corroborated by the WAVE-based investigation by Kortmann and Schröter (2017), who identify Asia as a region with strong areal features.

Mesthrie (2012: 785) confirms that, based on their WAVE profiles, “the regional set of Asian Englishes is also a typological set”, identifying similarities in all domains of morphosyntax. Specific attention is paid to the “pro-drop features” F42, F43 and F44, which exhibit a distinct agglomeration in the Asian-Pacific region (see the map in Mesthrie 2012: 787). Other features associated with pronoun omission, such as *it* in referential and non-referential “*it-is* constructions”, and subject relative pronouns, show strong attestation in the varieties discussed here (Table 4.3).

In the Asian subset, the exception is again low-proficiency Butler English in India, which commonly patterns with pidgins and creoles structurally (Mesthrie 2012: 802). Among the other six Asian varieties, Mesthrie identifies three subsets, Singapore and Malaysia, India and Hong Kong, and Pakistan and Sri Lanka, running somewhat against the likely expectation of a South vs Southeast Asian cluster, but reflecting the close historical and linguistic ties between Singapore and Malaysia (Mesthrie 2012: 804).

⁹ Philippine English is added in eWAVE 2.0 (2013), but not part of the analyses from Kortmann and Lunkenheimer (2012) discussed here.

Table 4.3: WAVE pronoun omission features in Asian varieties¹⁰

Feature	42	43	44	46	47	193
pro-drop	object	referential subject	dummy subject	referential <i>it is</i>	non-ref. <i>it is</i>	relative pronoun
Indian E	A	A	A	B	B	C
Hong Kong E	A	A	D	?	A	A
Singapore E	A	A	A	B	B	D
Malaysian E	A	A	A	D	B	D
Pakistani E	B	B	C	D	D	D
Sri Lankan E	B	B	B	D	D	D
Butler E	B	B	B	A	B	B

Regarding their overall morphosyntactic profiles, Mesthrie (2006) observes a “broad dichotomy” between different New Englishes of the *preserving* and the *deleting* type, found in Africa and Asia, respectively. These opposite tendencies can be traced back to the typological tendencies of the corresponding substrates, leading to “anti-deletion”, or preservation of synthetic structures, in one, and massive deletion of morphology, or increasing analyticity, in the other type of varieties (Mesthrie 2006: 466). Besides the pronoun-deletion features provided in Table 4.3, cases in point are BE deletion, lack of past marking, and article omission (see also Kortmann and Szmrecsanyi 2009 on simplification in contact varieties of English). However, identifying the mere presence or absence of a given feature is not always sufficiently informative. Sharma states that

[A] necessary next step in the identification of shared features is therefore to move from broad surface comparisons towards a close characterization of the grammatical conditioning of a given feature. This has been shown robustly in the variationist sociolinguistic literature to be a crucial element in establishing true parallelism between grammars and between speech communities. (Sharma 2012b: 215)

Percillier (2016) discusses the distribution and frequency of various structural features in Southeast Asian learner and indigenised varieties in Indonesia, Malaysia, and Singapore. He attests greater differences in syntax than in morphology and accent, with the highest amount of non-standard variants in the learner variety Indonesian English (Percillier 2016: 187). With regard to null subjects, the highest

¹⁰ The rating codes stand for A = pervasive or obligatory, B = neither pervasive nor extremely rare, C = feature exists but is extremely rare, D = attested absence of feature, ? = no information available.

number of zero pronouns is also attested for Indonesian English, followed by Malaysian and Singapore English (Percillier 2016: 92–93, but note that for the purpose of his investigation Percillier uses predominantly Singaporean data from ethnically Malay speakers, see also section 5.5.1).

To evaluate the status of morphosyntactic features in the three Asian Englishes discussed here, like the present investigation a number of comparative studies have used the respective ICE corpora as a database, e.g. Mukherjee and Gries (2009) on collocates in Asian Englishes. They relate Schneider's evolutionary stage to the amount of *collostructional nativisation* found in the respective varieties: the more advanced a variety is in the evolutionary cycle, the more dissimilar are its collostructional preferences to those attested in the British superstrate. Given the clear collocational association of null subjects to individual lexemes found in chapter 3 for British English, it is worth investigating how these patterns present in the Asian varieties.

Winkle (2015) is concerned with non-canonical sentence structures like left and right dislocation, fronting, existentials, *it*-clefts and pseudo-clefts in various spoken varieties of English, including the Asian Englishes, based on the respective spoken ICE components. While null subjects are not part of the analysis, the high amounts of left dislocation in Indian English can possibly be explained by transfer features of its topic-prominent substrates. This tendency is not reported to the same degree for the other Asian Englishes, a fact that is explained by the presence of overt topic markers in both Hindi and Malayalam (Winkle 2015: 216). This convergence thus leads to a different choice from the feature pool provided by the topic-prominent substrates compared to Singapore and Hong Kong English, where topic-marking is conveyed by word order only (see also section 4.1.2).

Two recent studies tackle the distribution of null subjects in selected Asian Englishes. The structural factors clause type, person and reference are examined by Schröter and Kortmann (2016) as possible determinants of null subjects in ten files each from the conversation sub-corpora of ICE-GB, ICE-HK and ICE-SG. Excluding null subjects in co-referential coordinations, omission rates amount to 1.7% in ICE-GB, 3.9% in ICE-HK and 6.4% in ICE-SG. Both Hong Kong and Singapore English show distinct substrate-induced constructions favouring null subjects, such as the *it is*-construction and relative clauses for Hong Kong, and *got*-existentials and questions for Singapore English. This is reflected in higher omission rates for subordinate clauses in Hong Kong English, and for questions in Singapore English. Among non-referential pronouns, all three varieties show a preference for generic null pronouns, compared to dummy pronouns and existentials, which are largely restricted to the Singlish *got*-existential construction.

Tamaredo and Fanego (2016) investigate the relation of morphological marking on third person singular verb forms and null subjects in ICE-SG and

ICE-IN, including texts from four different registers represented in ICE (five files each from the spoken informal, spoken formal, written informal and written formal sub-corpora, Tamaredo and Fanego 2016: 105). While subject omission is more common for Singapore English in general (approx. 9% vs 6% in Indian English), the difference is most pronounced when verb class is taken into account: of all verb classes, modal auxiliaries have the highest omission rate in Singapore English, while Indian English favours lexical verbs. Both varieties show a distinct decline in subject omission with non-modal auxiliaries (Tamaredo and Fanego 2016: 106). In the subset investigated, both varieties exhibit significantly more referential than non-referential null pronouns. Agreement shows a favouring effect for subject omission in the class of lexical verbs, but not for non-modal auxiliaries in both varieties. Moreover, the effect is too weak to override the strong tendency of the English superstrate for overt pronouns. The proposed explanatory factors of lexical frequency and collocational patterns are addressed in the analysis in section 5.2.2.

4.4.2 The comparative variationist method

This section provides methodological preliminaries underlying the comparative analysis in chapter 5. The variationist method of analysing grammatical variation was introduced in section 2.2, and successfully applied to null subjects in spoken British English in chapter 3. This method has also been used as a comparative measure to elucidate similarities and differences between varieties of English (e.g. Meyerhoff 2009, Poplack and Tagliamonte 2001, Tagliamonte 2008). The aim is to identify shared constraints in the distribution of variants in order to gain insight into historical relations between varieties. This includes a comparison of the respective variable context, as well as the statistical significance and range of factor groups (Tagliamonte 2008: 131). The initial spark for this comparative sociolinguistic investigation was the search for the origins and development of AAVE. Poplack and Tagliamonte (2001) aim to uncover historical sources of AAVE by analysing and contrasting linguistic constraints on different categories of verbal marking in dialects of AAVE.

Sharma and Rickford (2009) employ a similar method, comparing the constraint hierarchy for BE deletion across a range of contact Englishes, including AAVE, Indian English, and Singapore English with constraints found for learner English. They establish parallels with West African substrates for AAVE and creoles, but find contrasting patterns in L2 and learner Englishes, concluding that this divergence provides evidence for substrate rather than universal learner effects. This finding is obscured by less refined categorisations and emphasises

the need for detailed structural categories in contrastive analyses (Sharma and Rickford 2009: 85, see also Sharma 2009: 190).

Meyerhoff (2009) evaluates the degree of transfer from the Oceanic language Tamambo to the English-lexifier creole Bislama, testing constraints on subject- and object-pronoun deletion. The aim is to operationalise degrees of transfer for a more structured and measurable analytic outcome. The method involves

subjecting data from different groups of speakers to the same multivariate analysis. *The factor groups and factors within groups are held constant across the different populations, and the resulting Goldvarb weightings can be compared. Differences in which factor groups are returned as significant, and differences in the ranking of factors within those groups, are interpreted as diagnostics of fundamentally different underlying grammars of variation.* (Meyerhoff 2009: 303, emphasis mine)

The crucial methodological principle of this comparative procedure is to use the identical set of structural constraints, and test them for each of the languages or varieties under investigation. This is a strong argument for the detailed analysis of British English null subjects in the preceding chapters 2 and 3, since other appropriate studies providing the necessary information are not available so far. Degree of transfer is categorised by Meyerhoff into weak transfer, or *replication*, *strong transfer*, and *calquing*, representing diagnostic conditions based on multivariate analysis of the different data sets (Figure 4.1, adapted from Meyerhoff 2009: 304)

Diagnostic condition	Same significant factor groups	Same order of factor groups	Same order of factor groups and factors within groups
Degree of transfer	Weak transfer / replication	Strong transfer	Calquing

Figure 4.1: Meyerhoff’s taxonomy of transfer

In the present study, this comparative method is applied to the four varieties analysed, i.e. the superstrate and the three contact varieties. For comparison with the substrate language type, it translates into weighting the influence of superstrate, substrate, and universal factors in the different varieties, since analytical cate-

gories and factor levels from studies on Chinese are not directly comparable (see also section 4.1.3). Furthermore, the position constraint found for English null subjects is identified as a possible conflict site between English and the substrate languages, a condition that is also proposed as a measure for convergence of varieties of Spanish in contact with English by Torres Cacoullos and Travis (2010, 2015).

The degree of substrate influence and convergence in World Englishes is crucially influenced by their socio-historical background and ecology. The comparison of three varieties in similar linguistic ecologies with regard to subject omission, but at rather different stages of development, will try to further enlighten this issue. In the next chapter, Indian, Hong Kong and Singapore English are investigated for structural constraints on null subjects, following the procedure spelled out in chapter 3 and in the present section.

According to the description of the substrates and the processes of language contact, it is expected that

- Asian Englishes have a higher amount of non-referential null subjects than British English, based on their categorical absence in the substrates, and tendencies observed in contact languages;
- they exhibit a split between different persons, potentially related to third person marking;
- due to their presumed universal nature, factors like reference continuity and persistence are equally at work in contact varieties;
- due to their presumed English-specific nature, factors like clause type and position are less important for the contact varieties and present structural conflict sites;
- in contrast, the English-specific factor coreferential coordination combines the language specific preference with a context cognitively favouring referent accessibility, and is therefore not assumed to cause conflicts;
- tendencies for certain verb types, like a higher amount of especially first person overt pronouns with psychological verbs are universal, but collocational patterns of individual verbs are variety specific.

Overall, it is likely that Asian Englishes show a context generalisation for null pronouns, i.e. weakened constraints on their occurrence compared to Standard English, especially concerning the English-specific constraints to main clauses, initial position and coreferential coordination. This question is evaluated in chapter 6, based on the results of chapters 3 and 5.

5 Empirical comparison: Null subjects in Asian Englishes

So whatever English which is spoken here in India <,> I think these are all rule governed, because there is no language which is not rule governed <ICE-IN:028#31>

The preceding analysis in chapter 3 has shown that null subjects in Standard British English (ICE-GB) are more frequent than expected, but still rare from a cross-linguistic point of view. Language specific factor groups dominate the structural conditioning of variable subject pronoun realisation in ICE-GB. Chapter 4 has established how the relevant substrates in South and Southeast Asian varieties of English are different from British English regarding pronoun omission, and that influence in this domain of syntax, i.e. a higher frequency of null subjects, as well as a weakening of Standard English structural constraints, can reasonably be expected in the Asian Englishes. It is also assumed that substrate structures associated with topic prominence favour the emergence of parallel constructions containing null subjects. Universal and especially contact factors are likely to play a larger role in the contact varieties than English-specific factors.

This chapter tests these assumptions empirically by matching the constraints found for British English null subjects with each of the three Asian varieties individually. Results are discussed with regard to possible influence from their topic-prominent substrates. Section 5.1 shows how the structural factors and contexts of subject pronouns present differently in the Asian varieties compared to British English, parallel to sections 3.1 and 3.2. Difficulties that were encountered in the coding process of the Asian varieties specifically are addressed and illustrated with examples from the respective corpora. Section 5.2 presents an overview of the data structure and descriptive statistics of the Asian English corpora, followed by sections 5.3, 5.4, and 5.5 discussing the logistic regression models of Indian, Hong Kong, and Singapore English respectively. The comparison of the three Asian data sets is conducted in section 5.6, including a contrastive analysis of structural constraints found for Chinese null subjects.

5.1 Linguistic factors in the Asian Englishes

The extraction and annotation of pronominal and null subject tokens in the three Asian corpora in principle follows the same procedure and criteria outlined in sections 3.1 and 3.2 for the British data. However, as elaborated in section 4.3, the Asian varieties diverge structurally from the standard language enough as

to warrant a separate discussion how the linguistic factors investigated manifest in either of them. This section accounts for relevant distinctions, both in terms of concrete forms, as well as their frequency of occurrence of different linguistic contexts. Structural categories might be under- or overrepresented in the different corpora, and this possibly influences overall and category-specific drop rates. The relevance of linguistic factors for null subjects in the language type of the Asian substrates has been demonstrated by Jia and Bayley (2002) and Li et al. (2012) for Mandarin Chinese; demographic factors play but a minor role in the variable realisation of subject pronouns (see also section 4.1.3). This is corroborated by Schröter (2010), who finds that, unlike copula drop, the amount of subject drop in Singapore English is hardly affected by speech style and degree of formality. Among the three varieties discussed here, Singapore English especially is known to exhibit structures widely different from the English superstrate, and will consequently feature most prominently in this section.

5.1.1 The envelope of variation

Differences between the categorical contexts of individual varieties can be employed as a diagnostic of structural differences between grammatical systems (Tagliamonte 2006: 86, see also section 4.4.2). As elaborated in section 3.1, the comparison with the Asian varieties demands a less restrictive conception of the envelope of variation than for British English alone. To ensure consistent coding and comparable results, subject tokens still need to fit the definition provided for the analysis in chapter 3, i.e. overt or null pronominal subjects of finite verb phrases in non-formulaic utterances.

One example of divergent constructions is the occurrence of null subjects in a context usually invariant, the newly emerging Singlish discourse marker *know* (see also section 4.3.3). Wee (2003a) analyses the expression as an addition to the set of substrate induced invariant discourse particles in Singlish such as *lah*, *leh*, *mah*, *hor*, *what*, etc. (see e.g. Lim 2007 for an overview). Rather than a recurring syntactic process of pronoun deletion, the emergence of invariant *know* constitutes an instance of lexicalisation, albeit favoured by the tendency for null subjects in Singlish (Wee 2003a: 11). As in the case of British English, all discourse markers are excluded from the analysis of the Asian data; all instances of discourse marker *know* (bold in 5.1) are thus not considered an instance of pronoun drop as defined for the present study.

(5.1) But you see, not all husbands can do that Ø **know**. <ICE-SG:055#213:A>

Another structural phenomenon relevant for the variable context of subject pronouns is the frequent deletion of BE, both as auxiliary and copula in Asian Englishes (Table 5.1, see also Sharma and Rickford 2009). Again, it is Singlish that most radically omits BE across the different predicate types, and also allows for the omission of auxiliary HAVE. The differences in the WAVE ratings indicated in Table 4.3 for different kinds of pronoun omission are also reflected in the perceived frequency of BE deletion in the three ICE corpora.

Table 5.1: WAVE VP deletion features Asian Englishes

Feature #	174	175	176	177	178	179
Types of auxiliary deletion	auxiliary: progressive	auxiliary: <i>gonna</i>	copula: NP	copula: AdjP	copula: locative	auxiliary: HAVE
Indian E	B	B	C	D	C	C
Hong Kong E	A	A	B	B	B	D
Singapore E	A	B	A	A	A	A

BE deletion is encountered both in the form of overt subject pronouns without copula (5.2) and auxiliary (5.3), and as subject + BE deletion (5.4, line 2).

- (5.2) I Ø damn stupid *lah*. <ICE-SG:085#74:B>
- (5.3) B I'm not staring at you.
A Then who Ø you staring at? <ICE-SG:085#160:A>
- (5.4) 1 D She's not coming.
2 B Ø Ø Not coming *ah*?
3 D She Ø very fickle *leh*.
4 A ?Ø Just like you.
5 ?Ø Your friend *mah*.
6 Probably she Ø no mood to bathe *ah*. <GSSEC:062#1148–1151>

Copula deletion is especially common before modified adjectives (see also Ho and Platt 1993: 55). It is analysed as the transfer of Sinitic predicative adjectives, or property verbs (Ansaldi 2004: 135). The various instances in (5.4) show that cases of BE deletion in Singlish represent a continuum between such predicative adjectives and non-clausal or fragmentary utterances in spoken language. This goes beyond the scope of the present study, which is focused on the omission of subjects of finite verb phrases (for the exclusion of these and similar contexts, see also Nariyama 2004: 243, Torres Cacoulllos and Travis 2014: 21). Furthermore, (5.4) shows how blurred the boundaries are between clearly identifiable instances of

deletion (line 2 for subject + AUX, line 3 for copula BE), debatable (line 4), and highly unlikely (line 5) cases of an underlying subject + BE structure. Whether line 6 omits HAVE as a main verb or BE + PREPOSITION is also ambiguous. All instances of BE deletion, which are especially frequent in the Singaporean data, are thus excluded from the analysis (see also Schröter 2010 for the frequency of this phenomenon in different speech styles of Singapore English). For the present investigation, only line 1 as the sole complete subject + finite verb clause structure in this stretch of discourse is taken into account as an overt subject token in (5.4).

Concerning the linguistic constraints under investigation, from examination of the data differences between the three Asian corpora are most pronounced in the categories clause type, specific reference, and verb type, discussed in detail in the sections below. In accordance with the differences in the variable context, it is again Singlish that stands out with a number of idiosyncratic patterns.

5.1.2 Coreferential coordination

Based on comparable word numbers, the total amount of coreferential coordinations overtly marked with *and* is generally lower in the Asian varieties than in British English¹¹. The effect of coordination on subject realisation is commonly discussed as a condition specific to English (but see Torres Cacoullos and Travis 2015). Given the strong influence of this factor group on subject omission in Standard English (see section 3.6), the rarity of this construction in the Asian varieties possibly lowers their overall number of subject omissions by limiting a highly favourable context. However, the schema [NP_i VERB *and* Ø_i VERB] that forms the basis of the category under investigation here is only one specific form of coordination. Coordination can also be present in the form of lists and without overt conjunction; however, these instances of “asyndetic coordination” (Biber et al. 1999: 156) are not included as cases of coordination as defined here (5.5; Torres Cacoullos and Travis 2014: 26).

- (5.5) And daily now we_i read linguistic <,>
 Ø_i talk linguistic <,>
 Ø_i walk linguistic. <ICE-IN:010#51:C>

¹¹ Total number of coreferential coordinations marked with *and*: GB = 113, IN = 53, HK = 88, SG = 70; this difference is statistically significant in a Chi-square test with $\chi^2 = 35.99$, $df = 1$, $p < .01$.

On the other hand, lexically specific constructions, i.e. hendiadic [*I* GO_{1SGi} *and* Ø VERB_{1SGi}], and [*I* VERB_{1SGi} *and* Ø QUOTATIVE VERB_{1SGi}] have been found influential in US (Torres Cacoullos and Travis 2014) and British English (section 3.4.3). The aim is to establish whether these patterns are found in the Asian varieties as well and influence subject omission to a similar degree, in order to estimate how idiosyncratic these lexically specific constructions are across English varieties.

5.1.3 Clause type

The restriction of null subjects to main clauses is an English-specific constraint, and thus expected to be less strict in the Asian varieties. The two distinct types of subordinate clauses introduced in section 3.2.2, relative clause and conditional clause, are likely more relevant here than for British English due to substrate influence. Hong Kong English is described as more liberal concerning zero relatives (Gisborne 2000, 2009, see also section 4.3.2). This tendency is confirmed by the present data set. Zero relative pronouns in subject position are found in significant quantity in the Hong Kong data set, both for human (5.6) and non-human referents (5.7).

(5.6) And I have some friends Ø have visited there. <ICE-HK:021#59:A>

(5.7) Now how about the current issues Ø happened in Hong Kong.
<ICE-HK:022#170:A>

Singlish has a specific, substrate-induced conditional construction, termed *bare conditional* by Bao and Lye (2005, see also section 4.3.3). It is encoded without overt marking by conjunctions or prepositions (*if*, *in case of*, etc.) via juxtaposition of main clauses, leaving the interpretation of the conditional meaning to the hearer without explicit linguistic signposts. According to Bao and Lye (2005), this construction is part of a whole conspiracy of topic-prominent features transferred into the contact variety, and can be described as an instance of structural replication (see also section 4.2.1). Bare conditionals occur with both overt (marked bold in 5.8) and null subject pronouns (5.9).

(5.8) I was thinking
Ø **they** want to give me, give me me
Ø don't give me, done *loh* <GSSEC:022#42–43:B>

- (5.9) C Then you don't use any powder?
 B No but Ø put the powder
 Ø will become worse.
 '[If I] put on the powder, [it / the sweating] will become worse'
 <GSSEC:056#247–250>

Another replicated Sinitic construction in Singlish is the *X or not* question. Most likely transferred from Southern Chinese dialects like Teochew (see also section 4.3.3), this construction is used to form *yes-no* questions in spoken language, with and without overt subjects (5.10, 5.11). Within this construction, the lexically specific expression *can VERB or not?* is commonly mentioned impressionistically in descriptions of Singlish (5.12). All forms of *X or not* questions are indeed attested in the present data set.

- (5.10) You get me or not? <GSSEC-032#12:A>
 (5.11) Ø Remember or not? <ICE-SG:023#91:B>
 (5.12) Ø Can get raped or not? <ICE-SG:021#351:A>

As observed for British English (section 3.2.2), the distribution of persons across the different clause types is rather imbalanced, and shows recurring patterns determined by their communicative functions. Questions are most likely to contain second person referents as subjects, especially when compared to first person contexts, and null subjects in questions are also most likely to be omitted second person pronouns. This tendency is obviously affected by the communicative roles of different pronouns, so it can also be observed for Chinese, e.g. by Li et al. (2012: 105–106). In fact, in all varieties but Singlish it is exclusively second person zero pronouns that are attested in questions. Overall, the distribution of persons across clause types follows the same principles in all four corpora. Figure 5.1 shows the distribution of absolute token numbers of overt and null subjects combined in all four corpora. Concerning the overall amounts of different persons, Indian and Hong Kong English both display an imbalance towards first person pronouns. This is possibly influenced by the fact that many of the conversations represented in the corpora of these L2 varieties follow a less natural flow of discourse than those observed in Singapore and British English conversations.

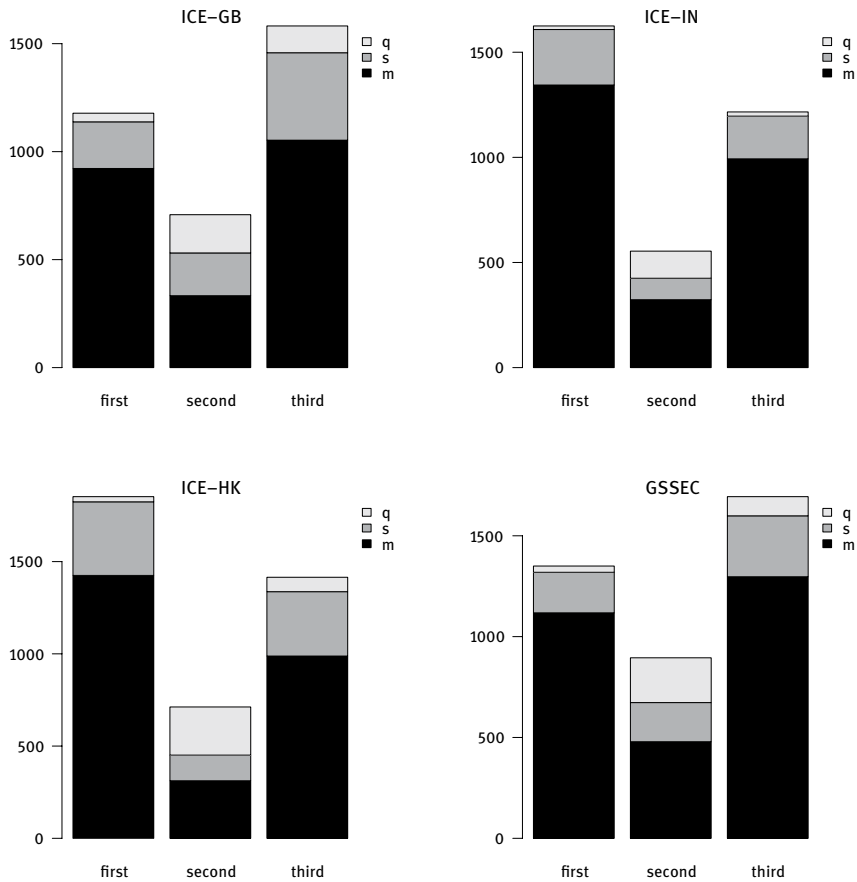


Figure 5.1: Comparison of absolute numbers of clause types by person

Given their status as non-native languages (see sections 4.3.1 and 4.3.2), the speech data in part resembles interviews, or stretches of reciprocal introductions, rather than naturally occurring conversations (5.13). The effect is most pronounced for the Indian data (see also Hansen 2018: 81–82).

- (5.13) A Let me introduce myself Dr Joseph <,>
B I'm [name] <,> from the depot of Karnatka <,>
A How long have you been teaching?
B No I am not a teacher I'm an officers steno working in a bank
<ICE-IN:026#1–6>

This also leads to the very low amount of questions in the Indian data outside of second person contexts. Pronouns in the GSSEC pattern very similarly to ICE-GB, with third person as the most frequent context, and a higher overall amount of second person than in both ICE-IN and ICE-HK.

For all varieties, main clauses form by far the most common environment for all persons. There do not appear to be significantly less subordinate clauses in the Asian varieties; questions, however, are underrepresented in ICE-HK, and especially ICE-IN.

5.1.4 Position

Like coreferential coordination, position is also an English-specific constraint on null subjects, and one of the most influential factor groups in the analysis in chapter 3 (see Table 3.5). The limitation to utterance initial position is strong enough to evoke frequent claims that it is the only possible context for Standard English null subjects outside coordination (e.g. Torres Cacoullos and Travis 2014, Weir 2012). This tendency has not been reported for other languages, and is thus expected to play a less central role in the Asian varieties. Common expressions preceding null pronouns, most notably in Indian English, are *and*, *then* and *but* (5.14). For the factor level “position 3”, the most common utterance initial expressions are *and then* (5.15), and introductory clauses like *I think*, *he says*, etc. (5.16).

(5.14) But Ø used to coach the children <ICE-IN:022#242:B>

(5.15) And then Ø make copy <ICE-HK:010#336:B>

(5.16) I think Ø need more <ICE-HK:050#470:A>

The position constraint on Standard English null subjects is thus a candidate for a structural conflict site in language contact (see also section 4.4.2, Torres Cacoullos and Travis 2015), and thereby a possible measure of degree of convergence in the contact varieties.

5.1.5 Person

As shown in section 3.4, British English exhibits clear differences for subject realisation rates between persons: while first and third person show similar drop rates, second person pronouns are by far least likely to be omitted. This is possibly explained by the different status and informational contribution of specific second person pronouns, which are usually focussed (Bailey 2011: 40). A subtype

of null subject languages, exhibiting a split system of subject omission between persons, is also represented by contact languages such as Bislama, and constitutes a possible configuration for the Asian contact varieties (see sections 2.1 and 4.2).

Another potential issue is the presence vs absence of morphological marking in English for different persons – only third person singular present tense forms are identifiable via morphology. However, third person singular contexts for subject pronouns still host a broad variety of possible forms (*he, she, it, this, that, there*, etc.) and referential status (referent in physical context, referent physically removed, syntactic pronouns, etc.); the contribution of this solitary morpheme for referent identification is thus limited at best. Token numbers for third person singular present tense verbs are rather small compared to base or past tense forms. Moreover, the Asian Englishes are known for the widespread omission of verbal morphology, including third person marking (e.g. WAVE feature F170). Systematic investigation of this factor is thus not conducted in the present study (but see Sato and Kim 2012 for a formalist approach, and Tamaredo and Fanego 2016 for a comparison of ICE-SG and ICE-IN).

5.1.6 Specific reference

In British English, non-referential, or non-specific reference only has a very weak favouring effect on null subjects. This category needs to be taken into account for the comparison here due to the substrate configuration of the Asian varieties, where non-referential subjects are not part of the inventory (see section 4.1). The absence of purely syntactic subject pronouns is also a strong tendency within contact languages and creoles (see section 4.2). It is thus expected that the Asian varieties will show a stronger preference for null subjects in non-referential contexts.

The share of overall non-referential contexts is similar in all four corpora (Figure 5.2). Apparently the Asian varieties do not avoid syntactic structures containing non-referential pronouns, like raising or extraposition, per se (but see Winkle 2015: 193–195 on the rarity of cleft constructions in the Asian Englishes). Indian English uses comparatively many generic pronouns, including generic usage of first person plural, as in (5.17), stating general advice rather than referring to a specific first person plural context, and third person plural, denoting unspecified inhabitants or the population in general of the city under discussion (5.18).

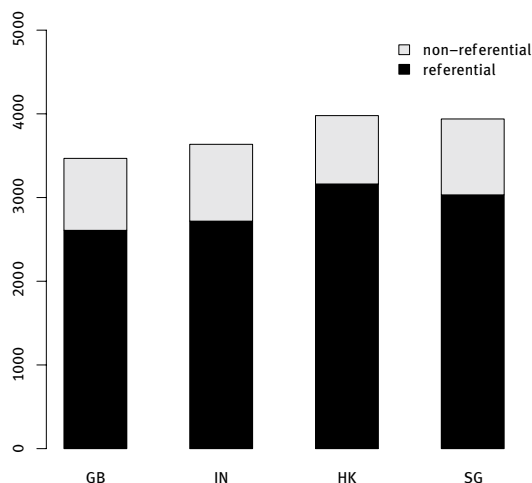


Figure 5.2: Referential vs non-referential pronouns

- (5.17) It's always better to read the poem or prose piece once again carefully before **we** give a talk about it. <ICE-IN:026#186:A>
- (5.18) In our place that is in Bhubaneswar <,>
first and foremost if you speak English
Ø means **they** will alienate you. <ICE-IN:010#110:B>

Existentials form another major subgroup of non-referential contexts. The substrate-induced *got*-construction is a highly productive way to express existential statements in Singlish: the files analysed in the GSSEC contain a total of 108 existential contexts; these are formed with *got* = 35, compared to *there* + BE = 66, + HAVE = 2, + USE = 2, + WILL = 3.

Referentiality is also a factor in the discussion of the *it is*-construction in Hong Kong English: while deletion of *it* is not attested for referential, it is pervasive for non-referential contexts, a contrast that is not indicated for Indian English or Singlish in WAVE (features F46 and F47). This contrast depending on referentiality is further proof for the grammatical, rather than phonological basis of this omission; this assumption is confirmed by Platt et al. (1984: 118), who state that the phenomenon is also observed in speakers who do not otherwise reduce consonant clusters.

5.1.7 Switch reference

While the effect of switch reference as a factor group is minor in British English, it is one of the most robust categories in cross-linguistic investigation, including both canonical and radical NSLs. The most common distinction within this factor group in the literature is binary, between referential switch and no switch, or reference maintenance. However, partial switch is a factor level found influential in studies on Chinese and contact languages (see sections 4.1 and 4.2), and thus possibly in the Asian varieties as well. Partial switch is found commonly in typical referential chains of introducing a referent in object position, a context categorised as switch reference in binary classifications. While the referent in partial switch contexts is not entirely continuous from the immediately preceding subject, its mention in another syntactic function constitutes an explicit linguistic introduction into the discourse context. Following Lambrecht's "principle of the separation of reference and role", this is in fact one major discourse function of objects (Lambrecht 1994: 184). In partial switch contexts, this referent, which is then both a familiar, and activated entity, is promoted to subject position in the following clause (5.19).

- (5.19) I wanted to kick him_i out of the house
 Ø_i Comes to my house <,>
 Ø_i stays there
 and then Ø_i says all that to me <ICE-IN:40#208:B>

The present study follows Jia and Bayley (2002), Li et al. (2012), and Meyerhoff (2000), and includes the third level partial switch to separate these contexts from those where the referent is not explicitly evoked in the preceding clause. This category also includes cases of partial referential overlap between subjects of neighbouring clauses, e.g. from first person plural to first person singular, but not cases where the surface form changes, but the referent remains constant, such as the shift from first person to second person in the case of speaker change; these are regarded as continuous reference (see also section 2.2.2).

5.1.8 Persistence

Persistence is supposedly a universal factor for subject realisation. The analysis in chapter 3 has shown that directly preceding null subject tokens trigger further null realisations, independent of co-reference, confirming observations by Torres Cacoullos and Travis (2014) and Wagner (2012, see section 3.4.3). This

effect is robust in British English, and is also expected to be relevant for the data here. Beyond the immediate context, the existence of clusters of null subjects is observed for ICE-GB (section 3.2.7), and is also attested for the Asian varieties, Singapore English especially. Third-wave sociolinguistic approaches interpret language variation as an indexical resource for stance taking and dynamic identity construction within communicative situations (Eckert 2008). This is confirmed by Leimgruber (2009, 2013), who finds clusters of different structural markers of colloquial Singapore English, in contrast with clusters of the respective standard variants, within the same stretches of speech. Following this assumption raises the question of the possible indexical meaning of null subjects in these clusters, and their possible co-occurrence with other salient localised features. However, Schröter (2010) finds no significant stylistic variation for null subject pronouns, making their use as sociolinguistic markers unlikely (see also section 5.5.4).

5.1.9 Verb phrase

The influence of semantic classes of verbs on subject realisation is found in various languages, whereas individual verb tokens might differ between languages and even varieties. The verb types investigated here, lexical verbs, psychological verbs, primary auxiliaries and modal auxiliaries, are not represented equally in the four corpora investigated. Asian Englishes in general feature more lexical verbs, which favour null subjects compared to the other verb types. The Singaporean conversations use less primary auxiliaries than the other three varieties, a quirk at least partly owed to the widespread use of copula drop in Singlish. High-frequency verb tokens feature prominently in co-occurrence patterns of null subjects in ICE-GB (section 3.3.3). High-frequency lexemes are remarkably similar across the different corpora, rendering lexical bias within this class of high-frequency tokens unlikely (an overview of the lexemes constituting 1% of all tokens in the different corpora is provided in Table B.15 in Appendix B).

The factor groups turn length and turn boundary do not contribute significantly to subject realisation in ICE-GB (see sections 3.4.1 and 3.5.2). The same is true for the Asian Englishes; turn length and turn boundary are thus no longer part of the following analyses. As a basis for the regression analysis, the next section provides an overview of the descriptive statistics for the three Asian data sets investigated.

5.2 Overview Asian Englishes: Descriptive statistics

This section provides an overview of the raw frequencies and descriptive statistics of the Asian English data sets and draws comparisons with ICE-GB as described in section 3.3, including the amount of null subjects across corpora, collocational preferences, and token counts for the different categories.

5.2.1 Different distribution of null subjects in the data sets

Parallel to the analysis of ICE-GB in chapter 3, twenty files from the different ICE-S1A sub-corpora are analysed, i.e. approximately 40,000 words, by about fifty speakers for each variety.

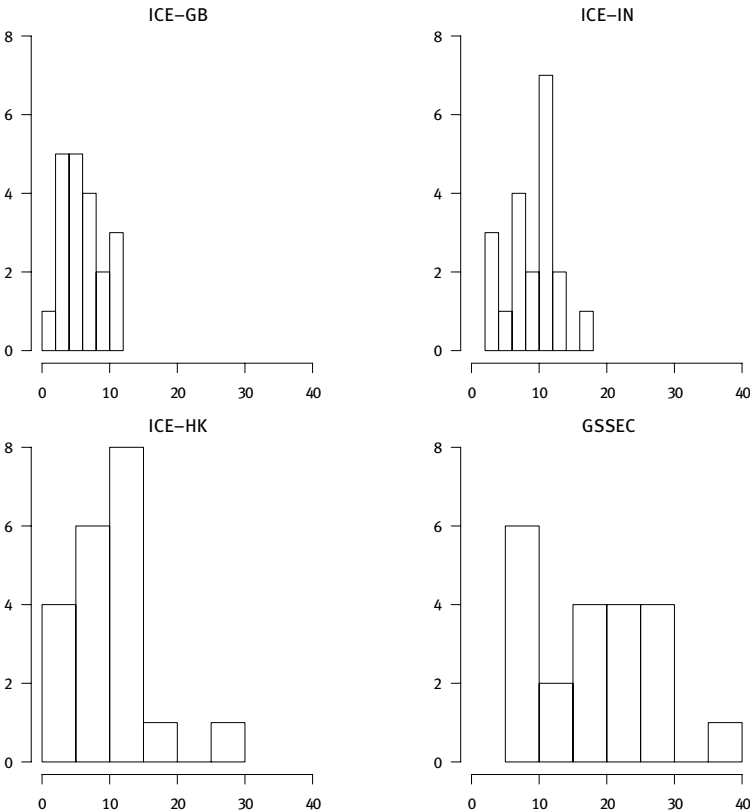


Figure 5.3: Comparison of null subjects per file

The data sets are fairly similar in that they amount to ca. 4,000 pronominal (overt and null) subject tokens per variety. All conversations include at least 3, and at most 40 null subject tokens per approximately 2,000 words, the targeted length of ICE files (Figure 5.3).

As for British English, deletion rates vary not only between conversations, but also between individual speakers. Tokens produced by speakers with categorical overt subject realisation are excluded from the following analysis (see also section 3.3.1). A first indicator for the differing propensity for null subjects in a variety is provided by the amount of categorical speakers in the four corpora (Figure 5.4): 24% in British, 17% in Indian, 8% in Hong Kong, and 4% of all speakers in Singapore English are invariant in their use of overt subject pronouns; null subjects are clearly most widespread within the speaker population in Hong Kong and Singapore.

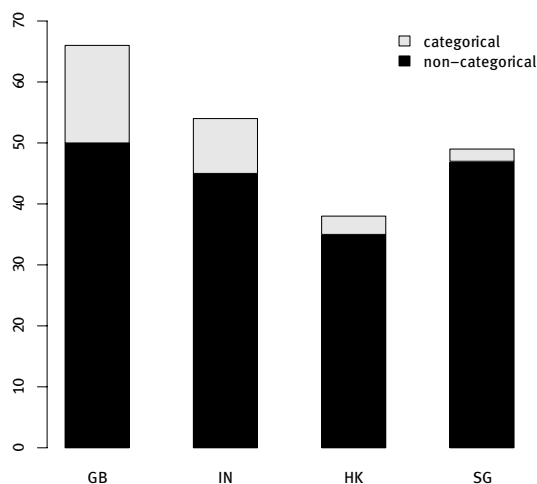


Figure 5.4: Amount of speakers with categorical overt subjects

Excluding speakers with categorical subject realisation yields the token numbers of pronominal subjects provided in Table 5.2 below, which form the basis for the following analyses (numbers for ICE-GB are reproduced here from section 3.3.1 for ease of comparison).

Table 5.2: Subject tokens all 4 corpora

	tokens total	null subjects	% null subjects
ICE-GB	3,468	130	3.8
ICE-IN	3,636	185	5.09
ICE-HK	3,979	206	5.18
GSSEC	3,939	468	11.88

Taking into account only contexts and speakers with variable realisation, the average deletion rates of roughly 5% for Indian and Hong Kong English hover surprisingly close to those found in ICE-GB. Besides a wider attestation within its speaker population, Singapore English also shows a distinctly more frequent occurrence of null pronouns. Different extra-linguistic factors are relevant for the different corpora; further details are discussed in the respective sections (5.3.1, 5.4.1 and 5.5.1). Figure 5.5 shows a comparison of the four data sets as percentage of zero subjects per conversation, including only the non-categorical speakers.

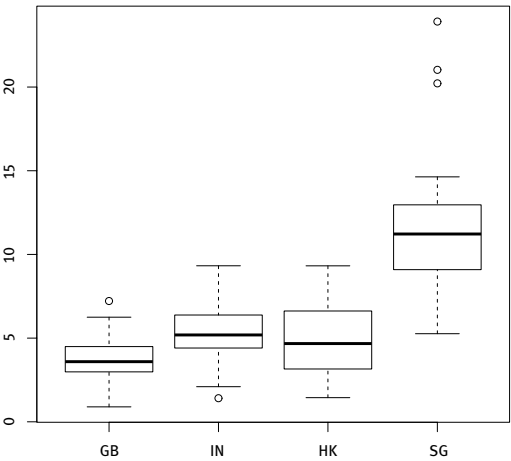


Figure 5.5: Comparison percentage of null subjects per conversation

This boxplot reveals some interesting observations about distributional differences within the four data sets. The British data is the most focused around the median deletion rate of 3.6%, with one outlier conversation slightly above the upper whisker. The Indian data is almost equally uniform, only one of the conversations constitutes an outlier with a deletion rate below the expected range.

The data from Hong Kong has no outliers at all, but the box representing the area between lower and upper quartiles is remarkably larger than for ICE-IN, indicating more internal variation. The Singapore data exhibits the widest distribution; both the box containing the majority of observations, and the whiskers indicating the conventional borders of measurements within the expected range, are more extensive than for the other data sets; moreover, three rather extreme outliers show a much higher null subject rate than the rest of the conversations in any of the other corpora. This broad overview of the four corpora already shows that major differences are not only to be expected between British and Asian Englishes, but also in Singlish compared to the other two Asian varieties.

Another relevant difference between the corpora is the share of coreferential conjunction within null subjects. While some studies (e.g. Wagner 2012) exclude this context altogether, it appears to make a substantial difference in the comparison here. Figure 5.6 shows how different the data sets behave in this respect.

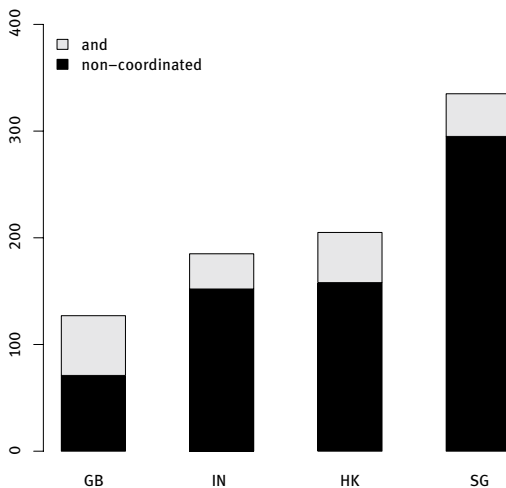


Figure 5.6: Comparison null subjects by coordination

Coreferential coordination is by far the dominant type of null subjects in ICE-GB, amounting to almost half of all zero pronominal tokens. The Asian varieties thus do not only contain less overt coordination constructions overall (see section 5.1.2), but the ratio of coordinations among null subjects is considerably smaller as well.

5.2.2 Collocation: Asian Englishes

The influence of individual verb tokens on subject realisation in British English is discussed in section 3.3.3. Two different aspects of co-occurrence patterns are addressed in chapter 3: the co-occurrence of both verb lemmas and specific verb forms with null subjects in general, and lexically specific patterns of coordination. Collocations are likely to be specific to individual varieties. Figure 5.7 shows the attraction and reliance scores of verb lemmas to null subjects in the Asian varieties (IN = orange, HK = purple, SG = blue; see Figure 3.7 for the British English equivalent).

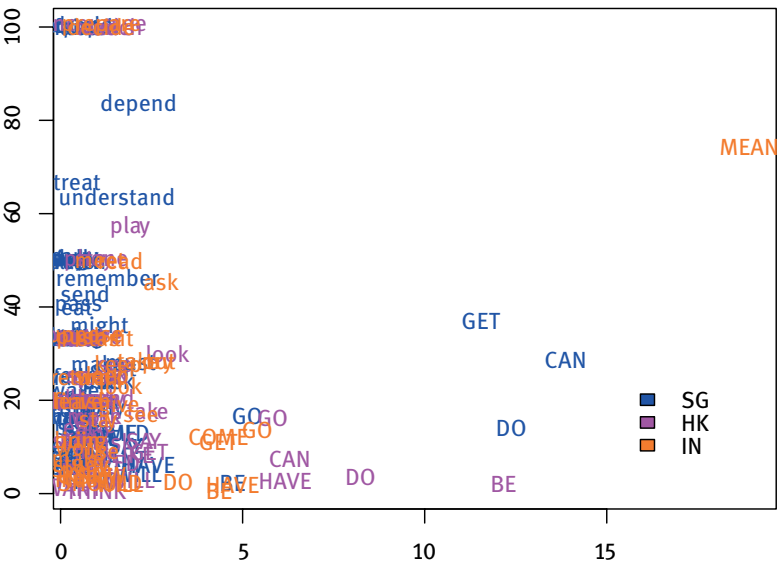


Figure 5.7: Collocates of null subjects Asian varieties

All varieties show large clusters of verbs at the lower end of the attraction scale, i.e. these lemmas represent only a small share of null subjects in each corpus. High-frequency items contribute the largest share of null subject tokens, and consequently show a tendency towards higher attraction scores. DO especially attracts null subjects in all varieties but Indian English. The lexeme GO yields comparatively high attraction and reliance scores, and behaves almost identical in all varieties.

Indian English generally exhibits comparatively low lexical affinities of null subjects to specific verbs, with one conspicuous exception, *MEAN*. *MEAN* in Indian English has both extremely high attraction and reliance scores, indicating that it constitutes a substantial amount of null subjects in ICE-IN overall, but also that a large share of the occurrences of the lexeme are found with null subjects, as in (5.20). This particular collocation is found in single instances in the other varieties, but not to the extraordinarily high degree as its attestation in Indian English.

- (5.20) If I'm absent Ø means everyone will look inquired about this
<ICE-IN:001#166:B>

Hong Kong English has high attraction, but low reliance scores for the primary auxiliaries. The high attraction of *BE* specifically to null subjects is unique to Hong Kong English compared to the other varieties, including British English. Almost all cases of null subjects with *BE* are part of either referential or, even more likely, non-referential *it is*-constructions (5.21).

- (5.21) Z It's just their way isn't it
A Ø is the Chinese character
<ICE-HK:030#888-889>

In Singapore English, the highest attraction score is achieved for the modal verb *CAN*. *DO* and *GET* are also highly attracted to null subjects, the lexeme *GET* specifically in the existential *got*-construction. *HAVE* is much less associated with null subjects in the GSSEC compared to the other data sets.

All corpora contain a handful of outliers in the form of low-frequency lexemes with a reliance score of 100, i.e. all instances of a specific lexeme occur with null subjects; none of these lexemes are shared between varieties, so these random patterns are not particularly informative. Apart from these outliers, Singapore English has a stronger tendency towards higher reliance scores in general, especially for the lexemes *DEPEND*, *TREAT*, *UNDERSTAND*, and *REMEMBER*. It seems the association of specific verbs with null subjects is stronger than in the other varieties; on the other hand, the class of verbs reliably attracting null subjects is larger in Singapore English.

The following tables present an overview of bidirectional collocational strength of verb forms with null subjects in either variety to evaluate the role of specific verb forms in the collocational patterns shown in Figure 5.7 above (using log transformed p-values of Fisher's exact test, the cut-off point is determined at $> 1.3 / < -1.3$ for attraction and repulsion respectively; for more details on the procedure, see section 3.3.3, Stefanowitsch and Gries 2005).

Table 5.3: Verb forms attracting null subjects ICE-IN

IN Verb Form	overt	zero	total	logpvF
<i>means</i>	6	35	41	41.322
<i>asked</i>	2	3	5	3.019
<i>sit</i>	2	3	5	3.019
<i>try</i>	2	3	5	3.019
<i>got</i>	31	7	38	2.778
<i>read</i>	3	3	6	2.733
<i>take</i>	6	3	9	2.155
<i>keep</i>	6	3	9	2.155
<i>puts</i>	2	2	4	1.906
<i>looked</i>	2	2	4	1.906
<i>see</i>	18	4	22	1.748
<i>ask</i>	3	2	5	1.697
<i>comes</i>	4	2	6	1.535
<i>go</i>	33	5	38	1.507
<i>went</i>	25	4	29	1.347

Table 5.4: Verb forms attracting null subjects ICE-HK

HK Verb Form	overt	zero	total	logpvF
<i>go</i>	34	12	46	5.79
<i>take</i>	12	5	17	2.939
<i>look</i>	10	4	14	2.38
<i>is</i>	235	23	258	2.144
<i>speak</i>	5	2	7	1.35
<i>try</i>	2	2	4	1.851
<i>stay</i>	2	2	4	1.851
<i>spend</i>	2	2	4	1.851
<i>read</i>	2	2	4	1.851
<i>see</i>	2	2	4	1.851

Table 5.5: Verb forms attracting null subjects GSSEC

SG Verb Form	overt	zero	total	logpvF
<i>got</i>	66	45	111	14.838
<i>depends</i>	2	12	14	9.472
<i>can</i>	117	47	164	9.146
<i>cannot</i>	19	17	36	7.074

Table 5.5: (continued)

SG Verb Form	overt	zero	total	logpvF
<i>don't</i>	200	54	254	5.461
<i>understand</i>	3	7	10	4.655
<i>go</i>	59	19	78	2.982
<i>must</i>	23	10	33	2.54
<i>haven't</i>	10	6	16	2.189
<i>remember</i>	7	5	12	2.11
<i>get</i>	25	9	34	1.917
<i>might</i>	9	5	14	1.792

The varieties show different patterns concerning the collocational strength of individual verb forms with null subjects. The most extreme outlier is found in Indian English (Table 5.3), *means*, with an extremely strong association to null subjects. The remaining collocates in ICE-IN consist predominantly of forms of motion verbs (*go*, *went*, *comes*) and quotative verbs (*asked*, *ask*). The latter are possibly also associated with the second position in coreferential coordination (see Figure 5.9).

The strongest collocate of null subjects in Hong Kong English is the verb form *go* (Table 5.4). Most instances are found in the second part of coreferential coordination, with or without *and* (5.22). Typically, these are part of biographic narratives, either of the speakers themselves, or of relatives, which present events in a linear order, or in the form of lists (5.23).

(5.22) He's born in Macau and then Ø just go to Hong Kong <ICE-HK:001#10:A>

(5.23) They finish studying
Ø go to work <ICE-HK:028#387:A>

Most remarkable about the Hong Kong collocates is the rather strong association of the verb form *is* with null subjects, considering the absence of any forms of BE amongst collocates of null subjects in the other varieties. With reference to Figure 5.7, it is likely the specific *it is*-construction attracting null subjects rather than the lexeme BE in general (5.21; see also section 4.4.1).

Singapore English has one extremely strong collocate, *got*, confirming the assumption that the high association of lexeme GET to null subjects is at least partly due to the existential *got*-construction (Table 5.5). However, *got* in the possessive sense of “obtained, received” also combines with null subjects, a collocational pattern that is found for ICE-IN and ICE-GB as well (5.24, Table 5.3, Table 3.2).

- (5.24) C Treat them very well but they got six months' bonus
 A My God, Ø got six months' bonus <GSSEC:013#422–423>

Singapore English is the only Asian variety at least partly sharing the preference of British English for null subjects with specific negated verb phrases (*haven't*, *don't* and especially *cannot*, 5.25).

- (5.25) A You can practice in your husband's car right?
 B Nicely caught, Ø catch you and Ø ban you
 A I mean your carpark *ah*
 B Ø Cannot.
 A Ø Cannot?
 B Ø Cannot. <GSSEC:062#362–372>

However, *cannot* is used rather idiosyncratically in Singapore English (see also section 5.5.4). Even more common is the co-occurrence of null subjects with the non-negated form *can* in the GSSEC (5.26).

- (5.26) B I_i also want to apply
 A Ø_i Can *ah*.
 I'm sure you_i can <GSSEC:012#34–36>

Must and *might* are further modal verbs prone to null subjects in Singlish. Overall, as hinted in Figure 5.7, Singlish has a set of verb forms that show relatively strong association with null subjects, especially compared to Indian and Hong Kong English, where, with the exception of single outliers, the p-values, and thus the collocational strength, for collocates of null subjects are significantly lower. The following tables show verb forms significantly repelling null subjects (i.e. with a log-transformed p-value < -1.3).

While lemmas and verb forms attracted to null subjects show clear differences and seem to be rather variety specific, repulsion of null subject concentrates on forms of the high-frequency primary auxiliaries BE and, to a degree, HAVE for all varieties (Table 5.6, Table 5.7). Singapore English additionally shows a strong aversion for null subjects with high-frequency psychological verbs like KNOW, THINK and MEAN (Table 5.8) – a stark contrast to Indian English, where the lexeme MEAN is the single most influential attractor for null subjects on the lexical level.

Table 5.6: Verb forms repelling null subjects ICE-IN

IN Verb Form	overt	zero	total	logpvF
's	230	0	230	-4.767
is	224	2	226	-2.73
am	107	0	107	-2.028
was	183	2	185	-1.935
are	270	5	275	-1.769
'm	95	0	95	-1.644
have	263	5	268	-1.63

Table 5.7: Verb forms repelling null subjects ICE-HK

HK Verb Form	overt	zero	total	logpvF
's	401	0	401	-9.157
are	179	0	179	-3.831
think	241	2	243	-3.376
'm	142	0	142	-2.912
've	68	0	68	-1.313

Table 5.8: Verb forms repelling null subjects GSSEC

SG Verb Form	overt	zero	total	logpvF
's	367	1	368	-18.529
'll	154	0	154	-8.116
was	174	6	180	-3.97
'm	106	2	108	-3.476
do	75	1	76	-2.784
know	50	0	50	-2.482
are	97	3	100	-2.421
're	48	0	48	-2.302
think	67	2	69	-1.688
have	103	5	108	-1.682
didn't	33	0	33	-1.561
mean	30	0	30	-1.374

Besides co-occurrence patterns of lemmas and verb forms with null subjects in general, Standard English exhibits lexically specific coordination patterns (see section 3.4.3.1, Figure 3.8 and Figure 3.9). Figure 5.8 and Figure 5.9 show the co-oc-

current patterns in the Asian data sets for first and second verb slot in coordinations, respectively.

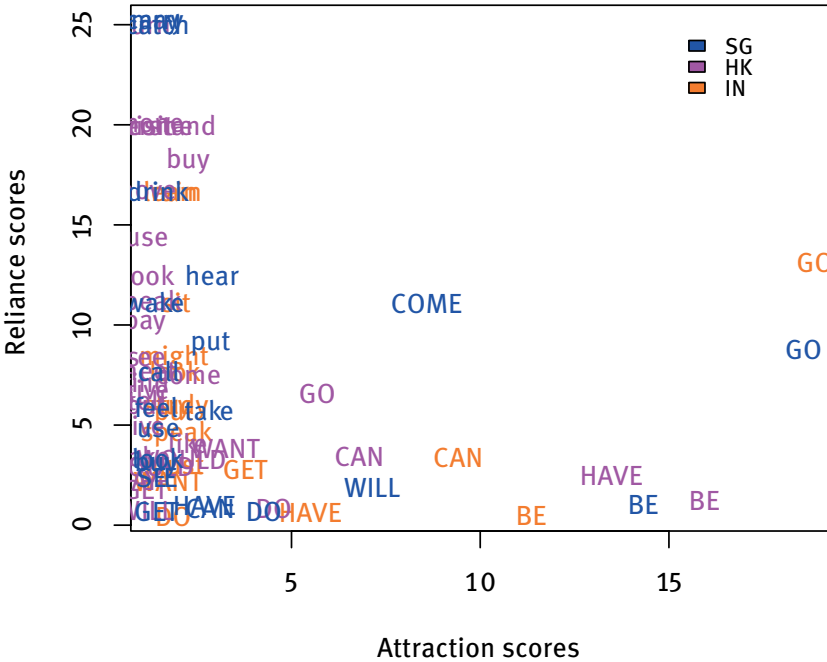
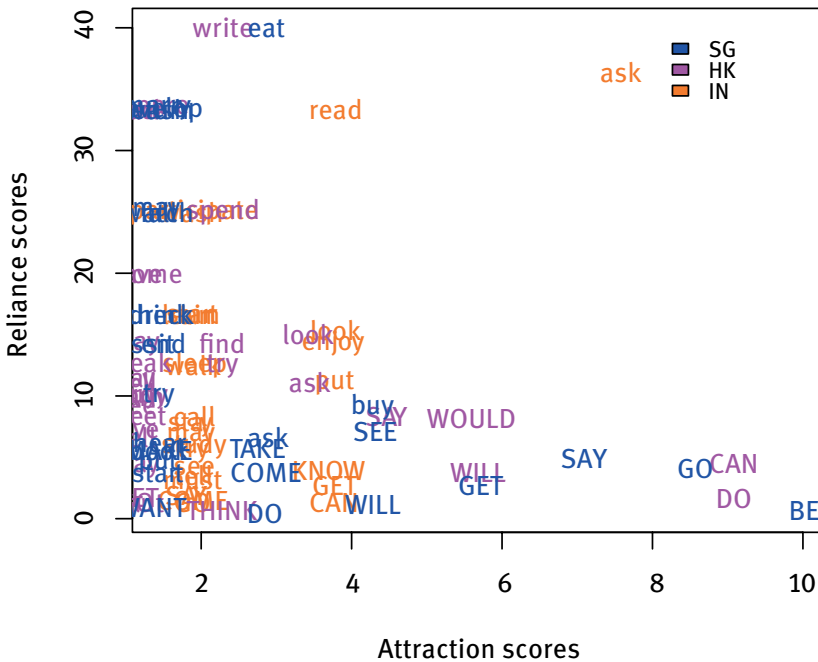


Figure 5.8: Attraction and reliance scores of verbs in position AND1 Asian Englishes

As observed for Standard English, the first verb slot in coordinations shows a strong tendency for motion verbs like GO and COME, especially in Singapore and Indian English. This association is even stronger than in ICE-GB, and shows that the occurrence of these prototypical hendiadic constructions is common across varieties and more representative of language-wide rather than variety-specific usage patterns. On the other hand, SEE and GET, strongly associated with the first coordinated VP in ICE-GB, are not attested here. Another diverging usage pattern is the high association of CAN with this syntactic slot in Indian and Hong Kong English.

Collocational patterns of the second coordinated slot are even more insightful for the distribution of null subjects, since this is the syntactic position most likely to trigger subject omission in Standard English. Similar to British English, forms of GO and GET are frequently encountered in this position, most notably in

Singapore English. SAY is the most common quotative verb, with the exception of Indian English, which favours ASK. However, for both positions the relatively low absolute number of overtly marked coordinations in the Asian Englishes limits the token numbers for this context, and consequently the more general validity of these observations. Overall, the co-occurrence patterns of individual lexemes with null subjects are idiosyncratic for the different varieties. An exception is found in the cases of GET, or more specifically, *got*, and forms of DO, which are encountered in several varieties investigated here. Verb forms disfavouring null subjects, on the other hand, are more likely to be shared across varieties.



egory for the three Asian varieties (for the respective numbers for ICE-GB, see section 3.4.4, Table 3.3).

Table 5.9: Linguistic factors and token numbers Asian Englishes

Factors		IN zero	IN overt	HK zero	HK overt	SG zero	SG overt
S total		185	3,451	206	3,773	468	3,471
English-specific	Coordination						
	n	152	3,431	159	3,732	422	3,447
	y	33	20	47	41	46	24
	Clause						
	main	175	2,485	168	2,557	386	2,508
	subordinate	7	803	34	853	36	661
	question	3	163	4	363	46	302
	Position						
	1	48	730	51	1,032	220	1,228
	2	24	704	28	1,104	75	799
	3	6	387	23	443	29	408
	x	107	1,630	104	1,194	144	1,036
Contact	Person						
	first	61	1,650	52	1,800	159	1,191
	second	18	566	28	684	102	793
	third	106	1,235	126	1,289	207	1,487
	Specific reference						
	ref	136	2,582	161	3,001	338	2,694
	non-ref	49	869	45	772	130	777
	Switch reference						
	switch	56	1,443	42	1,778	191	1,770
	partial	3	378	11	297	25	246
Universal	maintenance	126	1,630	153	1,698	252	1,455
	Persistence						
	pronoun	118	3,058	137	3,250	366	2,928
	NP	36	345	43	447	56	391
	zero	31	48	26	76	46	152
	Verb Type						
	lexical	139	831	102	841	178	920
	psychological	23	426	38	671	63	597
	auxiliary	17	1,622	45	1,742	123	1,385
	modal	6	572	21	519	104	569

The three sections below present the logistic regression models for each of the Asian varieties in turn. For details on the statistical method and model evalua-

tion, see sections 3.3 (descriptive statistics ICE-GB), 3.4 (full model ICE-GB), and 3.5 (first person ICE-GB); the following analyses adhere to the same guidelines. The analysis in chapter 3 shows that first person pronouns follow different constraints than the full set of subject pronouns; therefore, separate models for this context are computed for the Asian varieties. A comparative evaluation of the Asian varieties, including a weighting of the factor subgroups English-specific, contact and universal is provided in section 5.6.

5.3 ICE India

This section discusses further issues relevant for null subjects in Indian English (section 5.3.1), followed by the logistic regression model for the full Indian data set (section 5.3.2), and a separate analysis for first person pronouns (section 5.3.3).

5.3.1 Indian English: Extra-linguistic factors

Indian English is a collection of lects of varying proficiency and formality. The data in the ICE corpora represents the more educated formal speech variety, especially in L2 contexts (see also sections 4.3 and 4.4). Comparison of the null subject rate for the Indian data with a historically related variety reveals that it is similar to Mesthrie's findings for South African Indian English acrolect (Mesthrie 1992: 170).

ICE-India provides information on speaker gender, age, and L1 (see Hansen 2018 for a critical discussion). The data in the present study contains 26 female and 19 male speakers with variable subject usage. Figure 5.10 shows the distribution of null subject rates between female and male speakers.

Comparable to the age groups of ICE-GB (see section 3.3.2) the Indian data is divided into three age groups, 18–25 (9 speakers), 26–41 (26 speakers) and 42–50+ (10 speakers, Figure 5.11).

Both boxplots show that the distribution of null subject rates across these groups is rather uniform. Female and male speakers exhibit almost identical median deletion rates (on average, 5.9% zero for female, 5.2% for male speakers), and the female group only shows a slightly more diverse spread. There are no marked outliers in either group. The same is true for the different age groups. Older speakers show a slightly lower median deletion rate than the other age groups, but higher internal diversion. Chi-square tests confirm the lack of statistical significance for differences between the gender and age groups (with $\chi^2 = 0.06$, $df = 1$, $p = .81$ for gender, and $\chi^2 = 1.53$, $df = 2$, $p = .47$ for age).

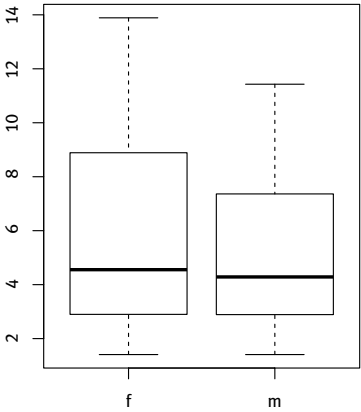


Figure 5.10: Percentage null subjects by gender ICE-IN

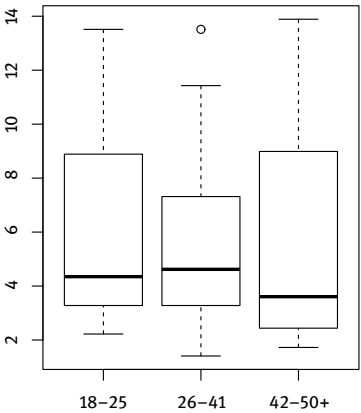


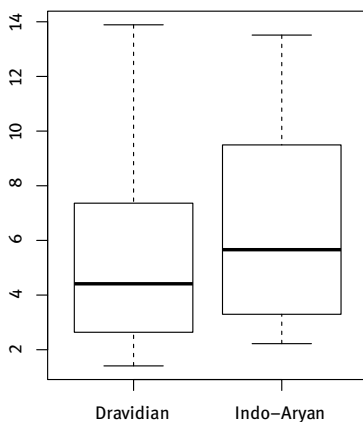
Figure 5.11: Percentage null subjects by age group ICE-IN

Another distinction is drawn between speakers of different substrate L1s. The present data mirrors the multilingual situation of the South-Asian subcontinent (Table 5.10, see also section 4.3.1). The 45 speakers provide the impressive number of thirteen different L1s. All of them indicate at least one additional language except English. Of the L1 languages, eight belong to the Indo-Aryan group, contributing 15 speakers, four to the Dravidian language family, contributing 28 speakers, and one to the Sino-Tibetan languages, contributing two speakers.

Table 5.10: L1 speakers ICE-IN

Language	Speakers	Language family
Bengali	3	Indo-Aryan (Eastern Zone)
Hindi	2	Indo-Aryan (Central Zone – Hindi)
Kashmiri	1	Indo-Aryan (Dardic)
Konkani	1	Indo-Aryan (Eastern Zone)
Marathi	4	Indo-Aryan (Southern)
Oriya	1	Indo-Aryan (Eastern)
Punjabi	1	Indo-Aryan (Greater Punjabi)
Urdu	2	Indo-Aryan (Central Zone – Hindi)
Kannada	9	Dravidian – Southern (Tamil-Kannada)
Malayalam	4	Dravidian – Southern (Tamil-Kannada)
Tamil	11	Dravidian – Southern (Tamil-Kannada)
Telugu	4	Dravidian – Southern (South-central)
Manipuri	2	Sino-Tibetan

Since there are not enough speakers of Sino-Tibetan languages in the data set to draw quantitative conclusions, the comparison is restricted to L1 speakers of the two major language families. While speakers of Indo-Aryan languages have a slightly higher null subject rate, Figure 5.12 shows that this difference is not significant; this is confirmed by a Chi-square test (with $\chi^2 = 0.41$, $df = 1$, $p = .52$).

**Figure 5.12:** Percentage null subjects by L1 ICE-IN

In conclusion, none of the extra-linguistic factors investigated for ICE-IN provides further insight into the frequency of null subjects in Indian English.

5.3.2 Indian English: Logistic regression model

This section presents the logistic regression model for the Indian English data. Its validity is evaluated following the same procedure as for British English in section 3.4. After a discussion of the full model, the influence of single factor levels is weighted. The Wald statistics for the factor groups in ICE-IN show that the factor groups position and specific reference are not statistically significant (Table 5.11).

Table 5.11: Wald statistics of factor group significance ICE-IN

Factor group	Chi-Square	d.f.	p
and	51.00	1	<.0001
Clause	33.47	2	<.0001
Position	5.55	3	0.1357
Person	19.70	2	0.0001
Specific reference	0.23	1	0.6316
Switch reference	13.44	2	0.0012
Persistence	79.52	2	<.0001
Verb type	138.09	3	<.0001
Total	307.98	16	<.0001

The remaining six factor groups have high statistical significance. A more parsimonious model for ICE-IN would thus omit these two factor groups; this minimal adequate model for ICE-IN is provided in Table B.16 in Appendix B. A comparison of the maximal and minimal adequate models shows that the reduced model does not perform significantly better than the full model (see Table B.17 in Appendix B). For the present purpose of evaluating the influence of the given structural factors on subject realisation by comparing the model fit, factor weights and constraint rankings across varieties, the two non-significant factor groups are kept in the model (see Meyerhoff 2009).

The logistic regression model is a good fit for the Indian data: the Model Likelihood Ratio Test (LRT) is significant with $p < .01$. A C value > 0.9 speaks for outstanding discrimination. The predictive power, indicated by the R^2 of 0.42, is also good. The test for overfitting reveals that optimism indexes are low (slope optimism < 0.05 , corrected C = 0.9, corrected $R^2 = 0.39$; the full output can be found in Table B.18 in Appendix B), the explanatory validity of the model beyond the given data set is thereby ensured. Multicollinearity is not an issue either: none of the estimated variance inflation factors exceeds 3 (Table 5.12).

Table 5.12: Estimated variance inflation factors ICE-IN

Factor level	vif
and: y	1.62
Clause: subordinate	1.43
Clause: question	1.22
Position: 2	1.35
Position: 3	1.16
Position: >3	1.54
Person: second	1.39
Person: third	2.02
Reference: non-referential	1.75
Switch: partial	1.18
Switch: maintenance	1.46
Persistence: NP	1.6
Persistence: zero	1.12
Verb type: psychological	1.08
Verb type: auxiliary	1.09
Verb type: modal	1.06

Table 5.13: Logistic regression model ICE-IN

	Coefficient	SE	Z	p-Value	
Intercept	-2.0652	2.0652	-7.729	<.0001	***
and: y	2.6157	0.3663	7.142	<.0001	***
Clause: subordinate	-2.473	0.428	-5.778	<.0001	***
Clause: question	-0.4143	0.6642	-0.624	0.5328	
Position: 2	-0.5817	0.2916	-1.994	0.0461	*
Position: 3	-0.7833	0.4654	-1.683	0.0923	.
Position: >3	-0.2694	0.2331	-1.207	0.2272	
Person: second	0.1685	0.3361	0.501	0.6162	
Person: third	1.0874	0.2554	4.257	<.0001	***
Reference: non-referential	-0.1251	0.261	-0.479	0.6316	
Switch: partial	-2.3005	0.6532	-3.522	0.0004	***
Switch: maintenance	-0.0034	0.2202	-0.015	0.9877	
Persistence: NP	1.3528	0.3095	4.370	<.0001	***
Persistence: zero	2.9449	0.3468	8.493	<2e-16	***
Verb type: psychological	-1.1495	0.2916	-3.942	<.0001	***
Verb type: auxiliary	-2.8534	0.2794	-10.212	<.0001	***
Verb type: modal	-2.9677	0.4553	-6.519	<.0001	***

Table 5.13 shows the full model for the Indian data, followed by a discussion of the influence of the different factor groups. Significance of individual factor levels

is indicated by asterisks. The discussion of the individual factor groups provides the Rbrul output of centred weights for the individual factor levels, and the range of the factor group as a whole. This output is based on the Rbrul and VarbRul standard sum coding, comparing the odds of a factor level with the mean of the whole factor group (Johnson 2009: 361, see also section 3.4). Among the statistically significant factor groups, coreferential coordination has a strong favouring effect on null realisation in Indian English (Table 5.14).

Table 5.14: Results for factor group coordination ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
y	1.308	53	0.623	0.787	57
n	-1.308	3,583	0.042	0.213	

In fact, the majority (33 of 53) of coreferential coordinations found in Indian English have zero rather than overt pronouns in the second verb phrase of the conjunction. Considering the lexical co-occurrence patterns discussed in section 5.2.2, which refer to coordinations both with overt and null subjects, it is remarkable that most of the coordinations with null subjects are indeed of the type [NP_i MOTION VERB + Ø_i VERB] (5.27), quite often even in combination with quotatives as the typical second constituent, i.e. [NP_i MOTION VERB + Ø_i QUOTATIVE] (5.28), while coordinations with overt subjects do not follow this pattern to the same degree.

- (5.27) This lady this Rheka she came and Ø told me <ICE-IN:010#183:C>
(5.28) But their parents whenever I beat them <,>
they came and Ø ask me <,> purposely why. <ICE-IN:085#167:A>

The factor group clause behaves similar to British English, but its effect is even stronger in Indian English. Subordinate clauses strongly disfavour zero, while the effect of the clause type question is neither statistically significant, nor decisive in terms of absolute numbers (Table 5.15).

Table 5.15: Results for factor group clause ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
main	0.962	2,660	0.066	0.724	54
question	0.548	166	0.018	0.634	
subordinate	-1.511	810	0.009	0.181	

In fact, overt subjects are almost categorical outside declarative main clauses, with rare exceptions in *that*-clauses (5.29).

- (5.29) It was supposed to start from today
 but the managing committee has asked <,>
 that Ø give some more time <ICE-IN:083#74:A>

The strongest effect within the factor group person is for third person pronouns. They have a strong favouring effect for zero compared to the reference level first person (Table 5.13). The factor weights show even more clearly that first person null subjects are avoided, and in fact less common than second person (Table 5.16).

Table 5.16: Results for factor group person ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
third	0.669	1,341	0.079	0.661	26
second	-0.250	584	0.031	0.438	
first	-0.419	1,711	0.036	0.397	

Unlike British English, second person has no such strong pronoun favouring effect; the factor weight is close to 0.5, and the factor level is thus not significant. In Indian English, first and second person pronouns behave more similarly, compared to the relatively high subject omission rate of third person pronouns. This split resembles the system found in creoles like Bislama, allowing for null subjects preferably in morphologically marked contexts. It is worth noting that by far the most favourable collocate for null subjects in ICE-IN, *means*, occurs in third person with null subjects exclusively, boosting omission rates for this context.

The universal factor group switch reference is statistically significant Indian English (Table 5.17).

Table 5.17: Results for factor group switch reference ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
switch	0.768	1,499	0.037	0.683	51
maintenance	0.765	1,756	0.072	0.682	
partial	-1.533	381	0.008	0.178	

However, its results are rather puzzling: both reference maintenance and full switch trigger null subjects, whereas partial switch has an extreme pronoun preserving effect. While this can be explained by possible ambiguity in partial switch contexts, a stronger pronoun preserving effect would still be expected for a full referential switch.

The two most decisive factor groups for ICE-IN are persistence (Table 5.18) and verb phrase; both of them are even more influential statistically for Indian English subject omission than the factor group coordination.

Table 5.18: Results for factor group persistence ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
zero	1.512	79	0.392	0.819	63
NP	-0.080	381	0.094	0.48	
pronoun	-1.433	3,176	0.037	0.193	

Preceding pronoun is a strong trigger for overt subject realisation. Both preceding lexical NP, and especially preceding zero favour null subjects. Zero to zero priming is thus noticeably strong in Indian English, as found in sequences like (5.30).

- (5.30) Ø Just rush up <,>
 Ø prepare everything <,>
 then Ø take our lunch <ICE-IN:030#152–153:A>

The interaction graph of the factor groups switch reference and persistence shows the blocking effect of partial switch on null subjects in all contexts, visible from the marginal amount of dots above the blue line for the respective columns. It also discloses the larger confidence interval, represented by the size of the grey bars, for this factor level compared to both full switch (s) and reference maintenance (m), which show less internal variability (Figure 5.13, see also section 3.4.3 for more information on interaction graphs). On the other hand, given the similar

likelihood of subject omission for both switch and maintenance, it is also obvious that this major distinction between full switch and reference continuity is not decisive for the significance of the factor group.

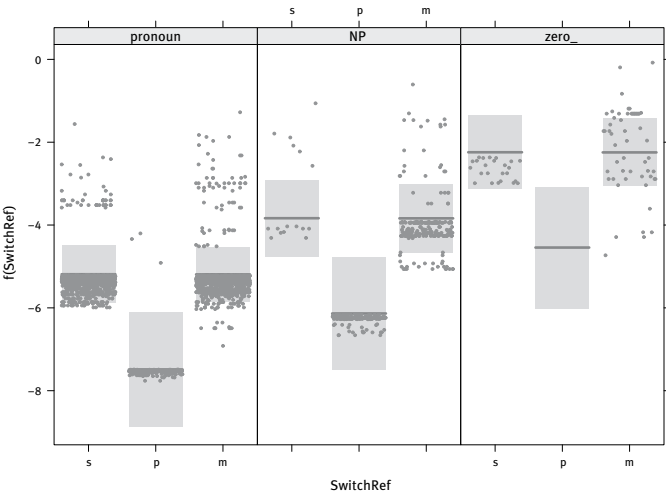


Figure 5.13: Interaction graph switch reference by persistence ICE-IN

Verb phrase is also a highly significant factor group for Indian English (Table 5.19). The disfavouring effects on null subjects of both modal, and especially primary auxiliaries, are prominent.

Table 5.19: Results for factor group verb type ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
lexical	1.743	970	0.143	0.851	62
psychological	0.593	449	0.051	0.644	
auxiliary	-1.111	1,639	0.010	0.248	
modal	-1.225	578	0.010	0.227	

Compared to lexical verbs, all other verb types favour overt pronoun expression. Indian English seems to be even more rigid than British English for this factor group. The interaction graph for the factor groups person and verb type reveals the strong association of third person pronouns with lexical verbs (Figure 5.14).

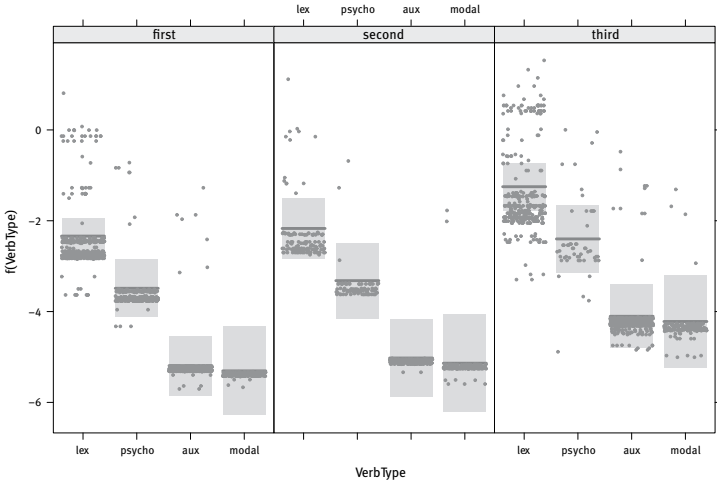


Figure 5.14: Interaction graph verb type by person ICE-IN

This is a partial explanation for the high omission rates of third person contexts. As for British English, auxiliary verbs do not omit pronouns in second person contexts at all. Additionally, first person zero in combination with modal verbs is not found in ICE-IN. Figure 5.14 also shows the patterning of first and second vs third person pronouns regarding the frequency of null pronouns, visible from the height of the lines indicating the median likelihood of omission.

As mentioned above, position is not statistically significant as a factor group. Concerning the different factor levels, it is remarkable that none of the positions have a strong pronoun favouring effect compared to initial position. This is evident from the factor weights that hover closely around 0.5, and the consequential low factor range (Table 5.20).

Table 5.20: Results for non-significant factor group position ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
1	0.409	778	0.062	0.601	19
>3	0.139	1,737	0.062	0.535	
2	-0.173	728	0.033	0.457	
3	-0.375	393	0.015	0.407	

The Indian English data also does not exhibit the same disinclination towards null subjects preceded by single items (position 2) as British English. The most

frequent expression preceding null subjects is initial *then*, often in linear narratives or enumerations as in (5.31).

- (5.31) Then Ø took over as principal of <,> a now college near Bijapur
<ICE-IN:026#10:A>

The factor range of referential status is even lower than for position (Table 5.21). The factor is clearly not relevant statistically, in spite of the topic-prominent configuration of the substrate(s) of Indian English.

Table 5.21: Results for non-significant factor group specific reference ICE-IN

factor	logodds	tokens	% zero	centred factor weight	range
referential	0.063	2,718	0.050	0.516	3
non-referential	-0.063	918	0.053	0.484	

The mosaic plot of referential vs non-referential pronouns shows that, apart from the higher overall token number of referential pronouns, they behave almost identical concerning the amount of null variants (Figure 5.15).

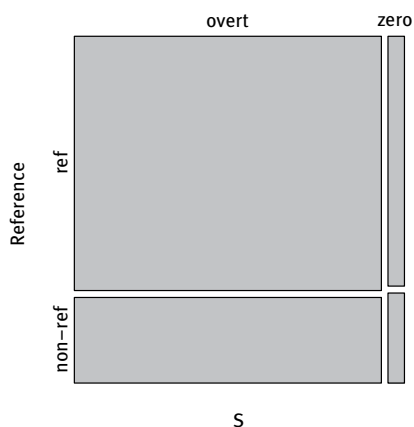


Figure 5.15: Referential vs non-referential pronouns ICE-IN

The most frequent case of non-referential null subjects in ICE-IN is found in connection with *means*, as in (5.32).

(5.32) Otherwise if we don't if I don't teach Ø means I will be feeling very sad.
<ICE-IN:022#281:B>

Only thirteen out of 49 non-referential null subjects in the Indian data occur outside this lexically specific context. This leaves non-referential null subjects, which are frequently described for vernacular Indian English, a rather marginal phenomenon for the educated variety represented here.

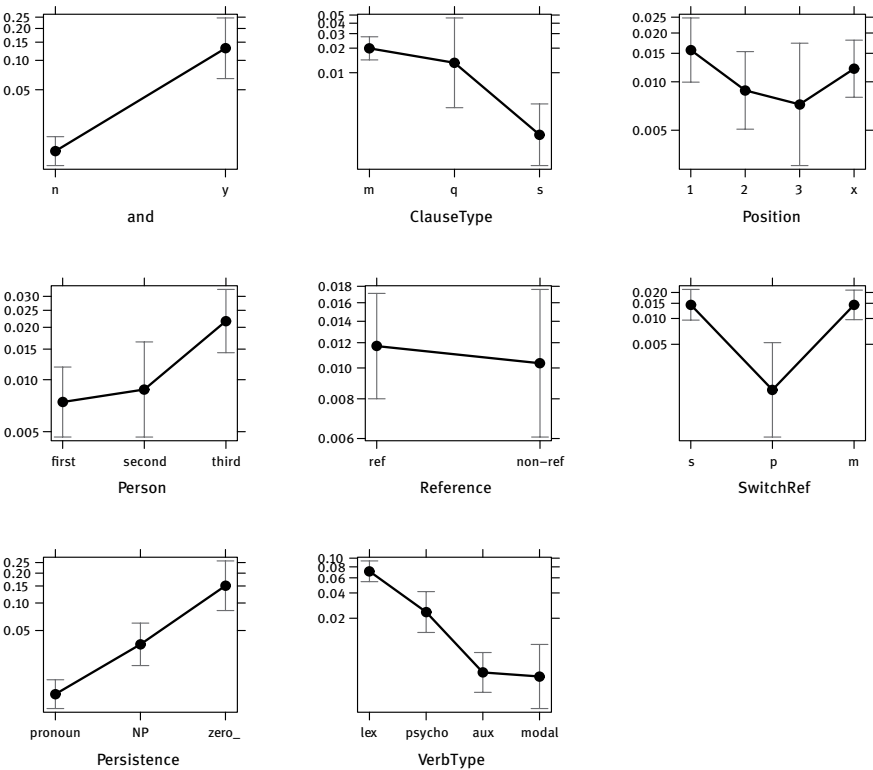


Figure 5.16: Predicted probabilities for factor levels ICE-IN

A summary of the direction of effects within the factor groups is provided in the effect plots for the full model for ICE-India (Figure 5.16). Factor levels are indicated on the x-axis, while the y-axis represents the predicted probabilities of null subjects in the respective context (see also section 3.4.4).

5.3.3 Indian English: First person contexts

The Indian data set contains 1,428 first person pronouns in variable contexts, with an omission rate of 4.27%. As for the British data, questions do not exhibit any null first person pronouns. A further invariant context for first person pronouns in the Indian data are modal verbs. Consequently, these factor levels are excluded from the analysis. These exclusions raise the omission rate considerably from first person contexts in the full set (3.5%).

In the statistical evaluation of the factor groups, coordination, switch reference, persistence and verb type remain significant, while clause and position are not. The resulting regression model (Table 5.22) is significant in the Model LRT ($p < .01$), and shows excellent discrimination ($C = .87$) and explanatory value ($R^2 = .38$).

Table 5.22: Logistic regression model first person ICE-IN

	Coefficient	SE	Z	p-Value	
Intercept	-2.5324	0.2151	-11.773	<.0001	***
and: y	2.0754	0.5139	4.038	<.0001	***
Switch: partial	-0.6056	1.1184	-0.542	0.5882	
Switch: maintenance	1.5361	0.5064	3.034	0.0024	**
Persistence: NP	2.5734	0.7445	3.456	0.0005	***
Persistence: zero	4.4727	0.5504	8.126	<.0001	***
Verb type: psychological	-0.9068	0.4466	-2.030	0.0423	*
Verb type: auxiliary	-2.1607	0.4788	-4.513	<.0001	***

All factor levels behave as expected, judging from the previous analysis: overt first person pronouns are strongly favoured by full, and even more by partial switch, and auxiliary verbs. As for the full model, persistence and verb type are the most influential factor groups. Zero to zero priming is even stronger for first person pronouns than for the full set in ICE-IN. There is no evidence, however, of a stronger pronoun favouring effect for psychological verbs in first person contexts compared to the full set. For Indian English null subjects, there are no clear differences in constraints between first person and the full data set, especially given the fact that the strongest constraints remain identical.

5.3.4 Short summary: Null subjects in Indian English

The analysis of the Indian data has shown that the extra-linguistic factors gender and age do not significantly influence subject realisation. Furthermore, no differences between L1 speakers of the two major language families, Indo-Aryan and Dravidian, are attested.

The regression model is a very good fit for the Indian data. Both the full data set as well as the subset of first person contexts can be explained well by the respective models, and adhere to similar constraints in their distribution.

While the English-specific constraints coordination and clause type are statistically significant, their role is less decisive than in Standard English. The strongest effects are measured for the universal factor groups verb type and persistence. The contact factors person but especially reference are more marginal in Indian English, and most likely concomitants of lexically determined preferences.

5.4 ICE Hong Kong

This section discusses further issues relevant for null subjects in Hong Kong English (section 5.4.1), followed by the logistic regression model for the full Hong Kong data set (section 5.4.2), and a separate analysis for first person pronouns (section 5.4.3).

5.4.1 Hong Kong English: Extra-linguistic factors

ICE Hong Kong includes information on speaker gender and age. While speaker age is provided, this category is not very useful for the subset of data under investigation here: in ICE-HK, the text type S1A (informal conversation) is dominated by young speakers, more specifically within the age group 21–25 (see also Hansen 2018: 90–91). As indicated by their occupation status, most of them are part of the student population. Of the 35 speakers contributing to the present data set, 31 fall into the age bracket 21–25; the remaining four speakers are divided between 26–30, 41–45 and 60+ years, which makes a quantitative comparison between age groups not feasible. At first sight, none of the speakers older than 25 shows a particularly unexpected deletion rate.

The data in the present study contains 24 female and 11 male speakers with variable subject usage. Figure 5.17 shows the distribution of null subject rates between female and male speakers.

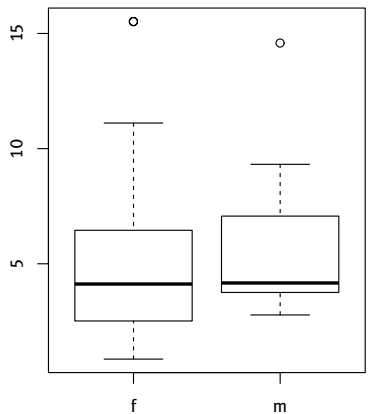


Figure 5.17: Percentage null subjects by gender ICE-HK

The internal dispersion is slightly higher for the female group of speakers; both groups contain one outlier, i.e. a speaker showing an exceptionally high rate of null subjects. As for the British and the Indian data, there is no measurable difference between genders regarding the average frequency of null subjects (on average, 5.3% for female, and 5.9% for male speakers; the Chi-square test is not significant with $\chi^2 = 3.42$, $df = 1$, $p = .06$). Given the changing proficiency rates in Hong Kong, an investigation including different age groups is desirable, potentially revealing differences between learner features in the language of older, and incipient nativisation in the language of younger speakers.

5.4.2 Hong Kong English: Logistic regression model

This section presents the logistic regression model for the Hong Kong English data. After a discussion of the full model, the influence of the single factor levels is weighted.

The Wald statistics for the factor groups show that, as for the Indian data, position and reference are not significant (Table 5.23). As for ICE-IN, they are kept in the model for comparative purpose; the minimal adequate model for ICE-HK is provided in Table B.19 and Table B.20 in Appendix B. The remaining six factor groups have very high statistical significance.

Table 5.23: Wald statistics of factor group significance ICE-HK

Factor group	Chi-Square	d.f.	p
and	71.28	1	<.0001
Clause	9.42	2	0.0090
Position	7.72	3	0.0521
Person	39.09	2	<.0001
Specific reference	3.36	1	0.0668
Switch reference	26.97	2	<.0001
Persistence	30.12	2	<.0001
Verb type	63.47	3	<.0001
Total	309.95	16	<.0001

Table 5.24: Estimated variance inflation factors ICE-HK

Factor level	vif
and: y	1.27
Clause: subordinate	1.18
Clause: question	1.12
Position: 2	1.37
Position: 3	1.35
Position. >3	1.82
Person: second	1.48
Person: third	1.74
Reference: non-referential	1.53
Switch: partial	1.43
Switch: maintenance	1.51
Persistence: NP	1.38
Persistence: zero	1.06
Verb type: psychological	1.18
Verb type: auxiliary	1.27
Verb type: modal	1.15

The logistic regression model is a good fit for the Hong Kong data: the Model LRT is significant with $p < .01$, a C value of 0.84 speaks for excellent discrimination. The predictive power, indicated by the R^2 of 0.28, is considerably lower than the value for the ICE-IN regression model. The test for overfitting reveals that optimism indexes are low (slope optimism $< .05$, corrected C = 0.83, corrected $R^2 = 0.26$; the full output can be found in Table B.21 in Appendix B), the explanatory validity of the model beyond the given data set is thereby ensured. Multicollinear-

ity is not an issue either: none of the estimated variance inflation factors exceeds 3 (Table 5.24).

Table 5.25 below shows the full model for the Hong Kong data, followed by a discussion of the influence of the different factor groups. Significance of factor levels is indicated by asterisks; factor groups marked non-significant in Table 5.23 are provided in square brackets.

Table 5.25: Logistic regression model ICE-HK

	Coefficient	SE	Z	p-Value	
Intercept	-3.7674	0.2828	-13.323	<.0001	***
and: y	2.4535	0.2906	8.443	<.0001	***
Clause: subordinate	-0.4462	0.2220	-2.010	0.0444	*
Clause: question	-1.3816	0.5509	-2.508	0.0122	*
[Position: 2]	-0.4562	0.2547	-1.791	0.0733	.
[Position: 3]	0.2573	0.2817	0.913	0.3611	
[Position: >3]	0.1680	0.2179	0.771	0.4407	
Person: second	0.6434	0.2741	2.347	0.0189	*
Person: third	1.3173	0.2123	6.204	<.0001	***
[Reference: non-referential]	0.4275	0.2332	1.833	0.0668	.
Switch: partial	0.2460	0.3891	0.632	0.5272	
Switch: maintenance	1.0271	0.2109	4.871	<.0001	***
Persistence: NP	0.6650	0.2323	2.863	0.0042	**
Persistence: zero	1.5970	0.3168	5.041	<.0001	***
Verb type: psychological	-0.5899	0.2285	-2.581	0.0099	**
Verb type: auxiliary	-1.5038	0.2059	-7.303	<.0001	***
Verb type: modal	-1.3748	0.2795	-4.919	<.0001	***

Among the statistically significant factor groups, coreferential coordination has a strong favouring effect on null realisation in Hong Kong English. It exhibits the highest statistical significance, and the broadest factor range (Table 5.26).

Table 5.26: Results for factor group coordination ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
y	1.227	88	0.534	0.773	55
n	-1.227	3,891	0.041	0.227	

In fact, a slight majority (47 of 88) of coreferential coordinations found in Hong Kong English have zero rather than overt pronoun in the second verb phrase of

the conjunction. The strongest collocational relation to this position is found for the modal verb *can* (5.33).

- (5.33) I want to be a doctor because they have a standard job
 and Ø can earn more money
 and Ø can save people's life.
 <ICE-HK:085#18:A>

The order of factors within the factor group clause is parallel to the ranking found for British English, but the overall effect of the factor group is not as strong in Hong Kong English (Table 5.27).

Table 5.27: Results for factor group clause ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
main	0.609	2,725	0.062	0.648	33
subordinate	0.163	887	0.038	0.541	
question	-0.772	367	0.011	0.316	

Questions have a clear pronoun preserving effect, while subordinate clauses show only a very weak effect towards overt pronouns. This can be explained by the marked propensity of Hong Kong English towards null subjects in relative clauses (see also sections 4.3.2 and 5.1.3). This higher likelihood of null subjects in subordinate clauses is not observed in the Indian data. The weakening of this superstrate-induced structural constraint in Hong Kong English is thus not necessarily a feature of contact or Asian varieties in general, but rather due to an idiosyncratic variety-specific shift in a certain type of subordinate clause. Outside relative clauses, Hong Kong English exhibits a number of null subjects in subordinate clauses introduced by *I think* (5.34), commonly in combination with the *it* is-construction (5.35).

- (5.34) I think Ø teach one class English <ICE-HK:084#327:A>
 (5.35) But I think Ø is quite crowd in Hong Kong University <ICE-HK:021#375:A>

The contact factor person is highly significant for Hong Kong English. The strongest effect is observed for third person pronouns, which strongly favour zero, especially compared to first person pronouns. This constitutes a deviance from the superstrate system, where first and third person are more similar concerning

their likelihood of omission, and a remarkable parallel to the equivalent observation on ICE-IN (Table 5.28).

Table 5.28: Results for factor group person ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
third	0.664	1,415	0.089	0.66	32
second	-0.010	712	0.039	0.497	
first	-0.654	1,852	0.028	0.342	

First person contexts have a strong pronoun favouring effect in Hong Kong English, they are least likely to permit null subjects. Like in Indian English, there is no distinct pronoun preserving effect found for second person contexts. Section 5.2.2 shows that the third person verb form *is* acts as an influential trigger for null subjects in ICE-HK. This collocate is possibly a partial explanation for the effect of the factor group person. Another plausible interpretation is the assumption of a creole-like split system of pronoun drop, found predominantly with the morphologically most transparent verb forms, a hypothesis that needs further investigation to be evaluated conclusively (see section 4.2).

The universal factor groups switch reference (Table 5.29) and persistence (Table 5.30) are statistically significant in Hong Kong English.

Table 5.29: Results for factor group switch reference ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
maintenance	0.603	1,851	0.083	0.646	25
partial	-0.178	308	0.036	0.456	
switch	-0.424	1,820	0.023	0.395	

The effects of the individual factor levels present as expected: reference continuity favours null subjects, while full switch favours overt pronouns. This is in accordance with cognitive principles of information management that predict less overt material for the encoding of more predictable referents (see section 2.1.2). Partial switch shows no influence in either direction, and is thus not significant as a factor level.

Table 5.30: Results for factor group persistence ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
zero	0.843	102	0.255	0.699	38
NP	-0.089	490	0.088	0.478	
pronoun	-0.754	3,387	0.040	0.32	

Preceding pronoun is a strong trigger for overt subject realisation. Both preceding lexical noun phrase, and especially preceding zero, favour null subjects. Figure 5.18 shows that the priming effect of immediately preceding null subjects is strongest in combination with continuous reference (m). This clear correlation is a marked contrast to both ICE-GB (Figure 3.13) and ICE-IN (Figure 5.13), where the omission-triggering effect of persistence is less dependent on reference continuity.

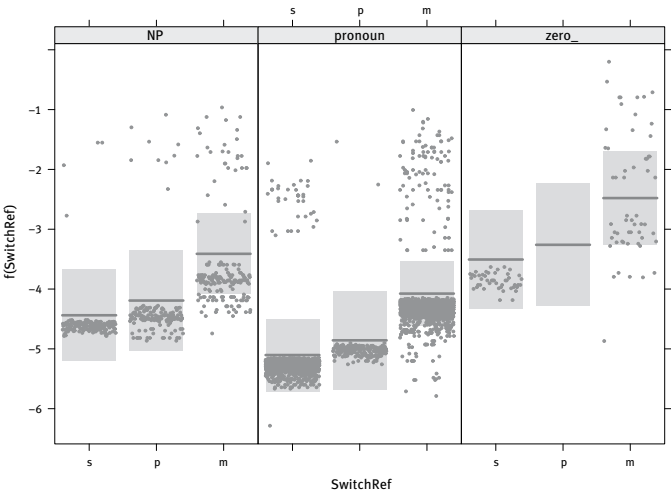


Figure 5.18: Interaction graph switch reference by persistence ICE-HK

As for Indian English, verb type is one of the most significant factor groups for Hong Kong English. Following the general tendency observed so far, null subjects are most likely with the least specific category lexical verbs. Compared to lexical verbs, all other verb types favour overt pronoun expression (Table 5.31).

Table 5.31: Results for factor group verb type ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
lexical	0.867	943	0.108	0.704	36
psychological	0.277	709	0.054	0.569	
modal	-0.508	540	0.039	0.376	
auxiliary	-0.637	1,787	0.025	0.346	

The most prominent effect is found for modal and primary auxiliaries, both triggering overt pronouns, although the effect is not as marked as in Indian English. As shown in Figure 5.19, null subjects with auxiliaries are predominantly found in third person contexts, and absent entirely in connection with second person pronouns. The dominance of third person zero in auxiliary contexts is based on the prevalence of subject omission in the *it is*-construction in Hong Kong English (see also section 4.4.1). The general preference for third person omission in Hong Kong English is visible for all verb types.

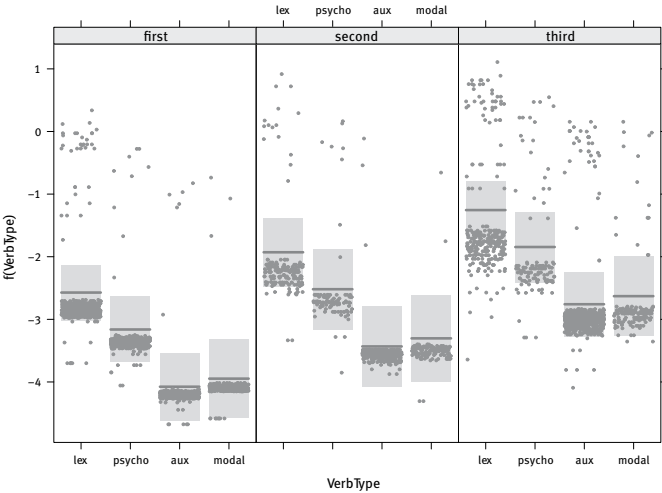


Figure 5.19: Interaction graph verb type by person ICE-HK

As mentioned above, position is not statistically significant as a factor group in Hong Kong English. Concerning the constraint ranking within this group, it is remarkable that initial position has no favouring zero effect at all. Like in ICE-GB, second position is most strongly triggering overt pronouns (Table 5.32).

Table 5.32: Results for non-significant factor group position ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
3	0.265	466	0.049	0.566	18
>3	0.176	1,298	0.080	0.544	
1	0.008	1,083	0.047	0.502	
2	-0.448	1,132	0.025	0.39	

Referential status is not relevant statistically (Table 5.33). Given the topic-prominent configuration of the local substrates, this is rather unexpected. Considering that of 45 non-referential zero pronouns, 18 are instances of non-referential *it is*, it is obvious that this type of null subject is marginal and not influential beyond this specific context in Hong Kong English.

Table 5.33: Results for non-significant factor group specific reference ICE-HK

factor	logodds	tokens	% zero	centred factor weight	range
non-referential	0.214	817	0.055	0.553	11
referential	-0.214	3,162	0.051	0.447	

Although specific reference is not significant as a factor group in isolation, following the WAVE ratings on referential and non-referential *it is*-constructions, its repercussion is expected to be visible in different behaviour of these two types (see section 4.3.2).

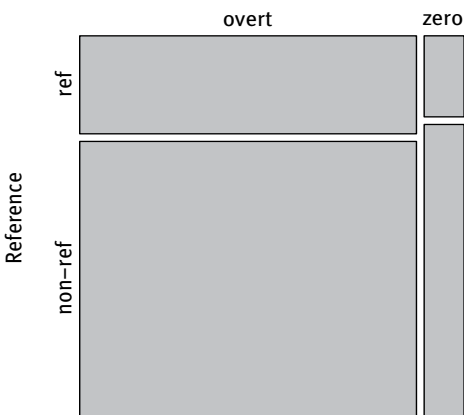


Figure 5.20: Referential vs non-referential *it is* ICE-HK

However, such an effect is not attested for the data investigated here. The mosaic plot shows that subject omission is equally common for referential and non-referential *it* is (Figure 5.20).

A summary of the direction of effects is provided in the effect plots for the full ICE-HK model (Figure 5.21).

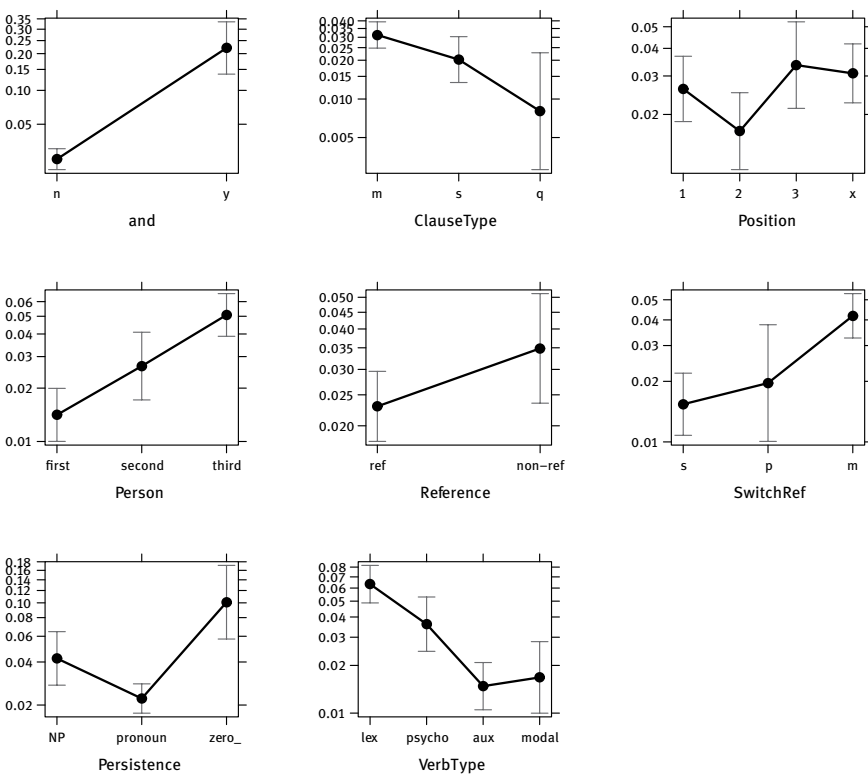


Figure 5.21: Predicted probabilities for factor levels ICE-HK

5.4.3 Hong Kong English: First person contexts

The Hong Kong data set contains 1,824 first person pronouns in variable contexts, with an omission rate of 2.85%. As obvious from the effects within the factor

group person (Figure 5.21 above), this rate is significantly lower than the null subject rate of 5.18% for all persons. As for the British and Indian data, questions in ICE-HK do not exhibit any null first person pronouns; all other relevant factor levels are included.

In the analysis of the remaining six factor groups, coordination, clause, switch reference, persistence and verb type remain significant, while position is not. The resulting regression model (Table 5.34) is significant in the model LRT ($p < .01$), and shows very good discrimination ($C = 0.83$) and explanatory value ($R^2 = 0.25$).

Table 5.34: Logistic regression model first person ICE-HK

	Coefficient	SE	Z	p-Value	
Intercept	-2.5794	0.2371	-10.877	<.0001	***
and: y	2.3157	0.4933	4.694	<.0001	***
Clause: subordinate	-0.8016	0.4625	-1.733	0.0831	.
Switch: partial	1.0094	0.6249	1.615	0.1062	
Switch: maintenance	1.6900	0.4979	3.394	<.0001	***
Persistence: NP	2.0340	0.5304	3.835	<.0001	***
Persistence: zero	2.8910	0.5183	5.578	<.0001	***
Verb type: psychological	-1.4426	0.4146	-3.479	0.0005	**
Verb type: auxiliary	-1.9150	0.4749	-4.032	<.0001	***
Verb type: modal	-1.8953	0.6308	-3.004	0.0027	**

As for the full model, coordination and persistence are the most influential factor groups. While the effect of persistence is stronger for first person, the direction of effects remains constant for both factor groups. Verb type continues to act as a significant predictor for the subset of first person pronouns. There is no clear evidence, however, of a distinct pronoun favouring effect for psychological verbs in first person contexts; although the pronoun-favouring effect is stronger than for the full model, this context remains the second most likely for first person pronoun omission as well, following lexical verbs, but permitting more null subjects than both primary and modal auxiliaries. For Hong Kong English null subjects, there are no clear differences in constraints between first person and the full data set. While the strength of effects varies for persistence and the factor level psychological verb, the significance of the factor groups, and the direction of effects remains constant.

5.4.4 Short summary: Null subjects in Hong Kong English

The analysis of the Hong Kong data has shown that the extra-linguistic factor gender does not significantly influence subject realisation. The distribution of speakers from different age groups in the present data set is not suitable for further analysis of this factor, but clarifies that the results obtained here are representative for a specific subset of Hong Kong English, the educated speech of a fairly young student population.

The regression model is a good fit for the Hong Kong data, although both the concordance index and especially the amount of variation explained by the model are lower than for Indian English, for both the full set and first person contexts.

The reported preference of Hong Kong English for null subjects in relative clauses can be confirmed for the present data, as well as the tendency for subject drop in both referential and non-referential *it is*-constructions. The strongest effects are measured for the factor groups coordination, verb type, and person, each one representing one of the subclasses English-specific, universal, and contact. The remaining English-specific factors, clause and position, have relatively little influence on subject omission, the substrate factor reference is even less significant. Concerning these subgroups of factors, there is thus no clear pattern for the origins of Hong Kong English null subjects.

5.5 Grammar of Spoken Singapore English Corpus

This section discusses extra-linguistic factors relevant for null subjects in Singapore English (section 5.5.1), followed by the logistic regression model for the full Singapore English data set (section 5.5.2), and a separate analysis for first person pronouns (section 5.5.3). The extensive structural divergence of Singapore English from its superstrate demands a complementary perspective on Singlish null subjects, provided by an analysis of various replicated substrate constructions and discourse-level features influencing subject pronoun realisation (section 5.5.4).

5.5.1 Singapore English: Extra-linguistic factors

ICE-Singapore does not provide any sociolinguistic speaker information. The analysis thus focusses on those files within the conversation sub-corpus of ICE-SG also contained in the Grammar of Spoken Singapore English Corpus (GSSEC), with supplemental GSSEC files to match the word number of the other

data sets (see also section 1.3). Concerning speaker age, given the provenance of the GSSEC data collection as part of a course assignment at university, the majority of the speakers consists of a relatively young student population, similar to the group sampled for ICE-HK (Lim and Foley 2004: 11–12). While there is no further detailed information on speaker gender and age available, the GSSEC contains information on speaker ethnicity (on the role of ethnic groups in the Singaporean speech community, see section 4.3.3).

The data in the present study contains speakers of all three major ethnic groups, 16 Chinese (C), 13 Malay (M), and 18 Indian (I). Figure 5.22 shows the distribution of null subject rates between speakers of the three ethnic groups.

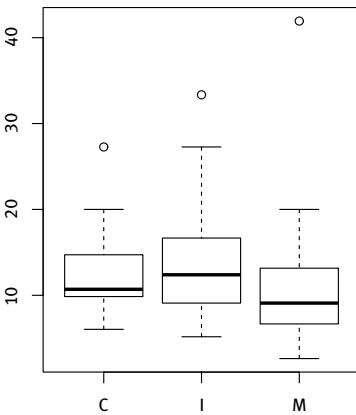


Figure 5.22: Percentage null subjects by ethnic group GSSEC

Malay speakers exhibit the lowest subject omission rates overall, Indian speakers the highest, with Chinese speakers in between. Each ethnic group contributes one outlier speaker with extraordinarily high amounts of null subjects. Furthermore, Chinese speakers behave most homogeneously compared to the other two groups. However, differences between these three groups are miniscule and not statistically significant. This is confirmed by a Chi-square test, and the observation that the factor levels show almost identical results in a logistic regression model with ethnicity as random intercept (the Chi-square test is not significant with $\chi^2 = 1.93$, $df = 2$, $p = .38$; a mixed-effects model with ethnic group as a random intercept is provided in Table B.22 in Appendix B). This is further substantiation of insights by Deterding and Poedjosoedarmo (1996) and Lim (2001) on the lack of grammatical differentiation between ethnic groups in Singapore English, which is rather found in pronunciation (see also section 4.3.3).

5.5.2 Singapore English: Logistic regression model

This section presents the logistic regression model for the Singapore English data. After a discussion of the full model, the influence of the single factor levels is weighted.

Table 5.35: Wald statistics of factor group significance GSSEC

Factor group	Chi-Square	d.f.	p
and	91.19	1	<.0001
Clause	28.66	2	<.0001
Position	32.43	3	<.0001
Person	5.10	2	0.0779
Specific reference	32.20	1	<.0001
Switch reference	7.46	2	0.0240
Persistence	27.74	2	<.0001
Verb type	45.08	3	<.0001
Total	248.25	16	<.0001

The Wald statistics for the factor groups show that, unlike in Indian and Hong Kong English, person is not a significant factor group for variable subject realisation in Singapore English (Table 5.35). As for ICE-IN and ICE-HK, the factor group is kept in the model for comparative purpose; the minimal adequate model for Singapore English is provided in Table B.23 and Table B.24 in Appendix B. The remaining seven factor groups have high statistical significance. The logistic regression model is not a particularly good fit for the Singlish data: while the Model LRT is significant with $p < .01$, a C value of 0.7 indicates barely “acceptable discrimination” (Levshina 2015: 259). The predictive power, indicated by the R^2 of 0.14, is very poor. Bootstrapping yields a slope optimism of .06, corrected $C = 0.7$, and a corrected $R^2 = 0.12$ (the full output can be found in Table B.25 in Appendix B). Multicollinearity is not an issue for the Singlish data set: none of the estimated variance inflation factors exceeds 3 (Table 5.36).

Table 5.36: Estimated variance inflation factors GSSEC

Factor level	vif
and: y	1.19
Clause: subordinate	1.18
Clause: question	1.23

Table 5.36: (continued)

Factor level	vif
Position: 2	1.16
Position: 3	1.14
Position: >3	1.44
Person: second	1.62
Person: third	1.77
Reference: non-referential	1.63
Switch: partial	1.2
Switch: maintenance	1.22
Persistence: NP	1.24
Persistence: zero	1.07
Verb type: psychological	1.25
Verb type: auxiliary	1.41
Verb type: modal	1.32

Table 5.37 below shows the full model for the Singlish data, followed by a discussion of the influence of the different factor groups. Significance of factor levels is indicated by asterisks; factor groups marked non-significant in Table 5.35 above are provided in square brackets.

Table 5.37: Logistic regression model GSSEC

	Coefficient	SE	Z	p-Value	
Intercept	-1.6706	0.1476	-11.316	<.0001	***
and: y	2.7616	0.2892	9.549	<.0001	***
Clause: subordinate	-0.6886	0.1971	-3.493	0.0005	****
Clause: question	0.6531	0.1910	3.418	0.0006	***
Position: 2	-0.6359	0.1466	-4.337	<.0001	***
Position: 3	-0.9595	0.2170	-4.422	<.0001	***
Position: >3	-0.4208	0.1389	-3.029	0.0025	**
[Person: second]	-0.3332	0.1588	-2.099	0.0358	*
[Person: third]	-0.2539	0.1389	-1.828	0.0676	.
Reference: non-referential	0.8381	0.1477	5.674	<.0001	***
Switch: partial	-0.0534	0.2413	-0.221	0.8248	
Switch: maintenance	0.2934	0.1143	2.567	0.0102	*
Persistence: NP	0.3976	0.1759	2.260	0.0238	*
Persistence: zero	0.9681	0.1925	5.028	<.0001	***
Verb type: psychological	-0.5643	0.1665	-3.389	0.0007	***
Verb type: auxiliary	-0.7828	0.1352	-5.789	<.0001	***
Verb type: modal	0.0047	0.1426	0.033	0.9734	

Except for person, all factor groups are statistically significant, and almost all factor levels achieve low p-values, i.e. statistical significance, as well. However, the relatively low log odds show that the influence of the individual factor levels is of limited weight. This underlines the importance of model evaluation, not just by assessing the significance of individual factor groups and levels, but also by checking measures of overall model validity, and the range of the different factor groups (see also section 3.3.5). Among the statistically significant factor groups, coordination has a strong favouring effect on null realisation (Table 5.38). In fact, the great majority (46 of 68) of coreferential coordinations found in Singlish have zero rather than overt pronouns in the second verb phrase of the conjunction. However, almost ten times as many null subjects (422) are still found outside this context, a stark difference to the distribution in ICE-GB, where 59 out of 130 null subject tokens are found in coordinated contexts (see also section 3.4.3).

Table 5.38: Results for factor group coordination GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
y	1.36	68	0.647	0.796	59
n	-1.36	3,695	0.111	0.204	

Lexical preferences for the overtly marked coordinating construction in the GSSEC are remarkably similar to ICE-GB, with the lexemes GO, SAY and GET as strong collocates for the second verb slot in coreferential coordination (5.36, see also Figure 5.9, Figure 3.9).

- (5.36) a. You start working then you go and Ø go for BBDC
<GSSEC:031#196:A>
- b. He put it on the heart and Ø say “what is what is that”
<GSSEC:081#516:B>
- c. He delivers and Ø get tips *what*
<GSSEC:056#119:A>

As for the other varieties in this study, the factor group clause is statistically significant for subject pronoun expression. However, in Singapore English the ranking of the factor levels is rather different (Table 5.39).

In Singapore English, questions have the strongest favouring effect for pronoun omission, while main clauses show only a very weak effect towards null subjects. Subordinate clauses clearly favour overt pronouns. Still, based on the overall higher amount of null subject pronouns, more diverse instances of null

pronouns in subordinate clauses are attested in the GSSEC compared to the other corpora: null pronouns in subordinate clauses can be coreferential either with overt (5.37), or null subjects of the main clause (5.38), or constitute a switch in reference (5.39).

Table 5.39: Results for factor group clause GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
question	0.660	343	0.131	0.659	32
main	0.017	2,741	0.136	0.504	
subordinate	-0.677	679	0.052	0.337	

- (5.37) Charles Dickens I read Christmas Carol that was that was cute
not too bad
so I_i think Ø_i can understand. <ICE-SG:090#402:B>
- (5.38) 1 B I_i may not want to go to Ministry of Labour for my attachment
2 A Why not
3 B Ø_i Don't know whether Ø_i got time.
4 A Why Ø_i got no time?
5 You_i still got a lot of time <ICE-SG:032#185–190>
- (5.39) So this is [a new doctor]_i
She_j went there
Wah, [cute guy]_i and all that
Then Ø_j found out
Ø_i was Richard's son. <GSSEC:013#723–726:C>

Lines 4 and 5 in (5.38) show that in this case *got* is not used as existential with dummy null pronoun, but as possessive with omitted first person subject pronoun, coreferential with the overt pronoun in line 1, and the omitted subject of the main clause in line 3. The strong tendency of questions towards null subjects can be explained by the widespread use of question formation strategies replicated from Chinese varieties like the *X or not* question (5.40, see also section 4.3.3), but also *wh*-in-situ questions without overt subject (5.41). As evident from (5.41 b.), Chinese question words occupy the place of the NP without movement. For *wèi shénme* ('why'), the placement depends on the focus of the question (compare 5.41 c and d).

- (5.40) Ø understand or not? <GSSEC:112#88:A>

- (5.41) a. *Zhāng zài shāng mǎi le liǎng jiàn chǎng chènshān.*
 Zhang in market buy LE two CL shirt
 'Zhang bought two shirts in the market.'
- b. *Zhāng zài nǎ mǎi le liǎng jiàn chènshān?*
 Zhang where buy LE two CL shirt
 'Where did Zhang buy two shirts?'
- c. *Wèi zhāng zài shāng mǎi le liǎng jiàn shénme chǎng chènshān?*
 why Zhang in market buy LE two CL shirt
 'Why did Zhang buy two shirts in the market?'
- d. *Zhāng wèi zài shāng mǎi le liǎng jiàn shénme chǎng chènshān?*
 Zhang why in market buy LE two CL shirt
 'Why did Zhang buy two shirts in the market?'
- (adapted from Yip and Rimmington 2004: 353–354)

- (5.42) Last time Ø go where *ah?* <GSSEC:058#440:B>
 Why Ø depends on results? <GSSEC:031#357:A>

Like in Hong Kong English, the weakening of the superstrate-induced structural constraint on the clause types permitting null subjects is at least partly due to a variety-specific shift in a certain type of clause, in the case of Singlish, questions. On the other hand, the relatively common occurrence of null subjects in subordinate clauses independent of referential continuity shows that Singapore English undermines this strong superstrate constraint more fundamentally than for example Hong Kong English, where the deviance is more restricted structurally, and in terms of frequency.

The third English-specific constraint, position, is significant for Singapore English subject realisation: as in ICE-GB, all positions except initial favour overt pronouns (Table 5.40).

Table 5.40: Results for factor group position GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
1	0.529	1,379	0.157	0.629	25
>3	0.082	1,122	0.122	0.521	
2	-0.118	831	0.088	0.47	
3	-0.493	431	0.065	0.379	

Initial null subjects in Singlish are frequently found in rather conventionalised expressions like *seems*, *depends*, *don't know*, but also with *can* / *cannot*. The latter three are commonly used as complete utterances.

The substrate-inspired factor group specific reference is clearly significant for Singapore English, with non-referential pronouns showing a stronger tendency for omission (Table 5.41).

Table 5.41: Results for factor group specific reference GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
non-referential	0.42	849	0.143	0.604	21
referential	-0.42	2,914	0.114	0.396	

Got is used regularly as existential marker, usually without overt non-referential subject (5.43). As observed by (Leimgruber 2009: 59), the mixed construction *there* + *got* is attested as well. It is not especially frequent, however, and tends to carry locative meaning (5.44).

- (5.43) Maybe they make one extra class on that day
 Ah, ok Ø got one extra class ah July <GSSEC:031#271–272>
- (5.44) Oh ya next time I buy myself in Boston
 There got shoes and everything <ICE-SG:021#253–254:A>

In Leimgruber's data, the use of existential *got* is a clear marker of informal language use; it amounts to only 5% of all existentials in the formal, but to 32% of existentials in the informal speech situation (Leimgruber 2013: 78), the same amount is found in the present data set (see also section 5.1.6). Existential *got* is another example of a construction-specific weakening of a more general structural constraint predetermined by the superstrate configuration.

While the universal factor switch reference is influential in Indian and Hong Kong English, it only shows limited influence on null subjects in Singapore English (Table 5.42).

Table 5.42: Results for factor group switch reference GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
maintenance	0.244	1,629	0.151	0.561	10
switch	-0.078	1,876	0.098	0.48	
partial	-0.166	258	0.093	0.459	

The direction of effects within this factor group mirrors the observations made in section 4.1.3 on Chinese: reference continuity slightly favours null subjects, a full referential switch has a weak, partial switch a slightly stronger pronoun favouring effect.

The effect of persistence is also a lot weaker in Singapore than in Hong Kong and especially Indian English. Confirming the direction of effect found for ICE-HK and ICE-IN, both preceding zero and preceding NP favour null pronouns, compared to the reference level preceding pronoun. A more detailed discussion of clusters of null subjects and their relation to persistence is provided in section 5.5.4 below.

Table 5.43: Results for factor group persistence GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
zero	0.491	197	0.234	0.62	23
NP	-0.028	429	0.131	0.493	
pronoun	-0.464	3,137	0.112	0.386	

The effect of verb phrase presents rather different in Singlish from the other varieties (Table 5.44). Psychological and especially primary auxiliary verbs strongly disfavour pronoun omission, albeit to a lesser degree than in the other Asian Englishes. The weaker effect of auxiliary verbs on overt subject expression is possibly related to the widespread use of zero copula in Singlish, often in combination with subject omission (see also section 5.5.4 below).

Table 5.44: Results for factor group verb type GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
modal	0.355	641	0.156	0.588	19
lexical	0.319	1,069	0.161	0.579	
psychological	-0.269	624	0.095	0.433	
auxiliary	-0.406	1,429	0.086	0.4	

The highest deviance from the other varieties in this study is found for modal verbs, which show the strongest tendency of all verb types towards null subjects in Singlish (see also Table 5.8). Examples include cases in utterance initial position familiar from British English (5.45), but also more idiosyncratic expressions (5.46). (5.46 a) and (5.46 b) show non-initial null subjects, both in combination with null objects of the lexical verbs, while the null subject in (5.46 c) occurs in a subordinate clause, and is coreferential with the null subject of the introductory main clause.

- (5.45) a. \emptyset can't be bothered <GSSEC:051#305:a>
 b. \emptyset should be no problem <GSSEC:077#55:a>
 (5.46) a. Yah *lor*, later \emptyset can buy \emptyset *lor* <GSSEC:051#305:a>
 b. No later \emptyset must finish \emptyset man <GSSEC:011#329:b>
 c. \emptyset don't know whether \emptyset should call my cousin
 <GSSEC:058#195:a>

As mentioned above, person is not statistically significant as a factor group (Table 5.45). In stark contrast to Indian and Hong Kong English, first person contexts are most favourable for Singapore English null subjects; the effect of the factor level is, however, barely significant.

Table 5.45: Results for non-significant factor group person GSSEC

factor	logodds	tokens	% zero	centred factor weight	range
first	0.217	1,287	0.122	0.554	9
third	-0.064	1,593	0.124	0.484	
second	-0.154	883	0.113	0.462	

The interaction graph of the factor groups person and verb type (Figure 5.23) shows that, especially compared to other varieties (Figure 5.14, Figure 5.19), there

is less obvious correlation between persons and specific verb types, both for overt and null pronouns.

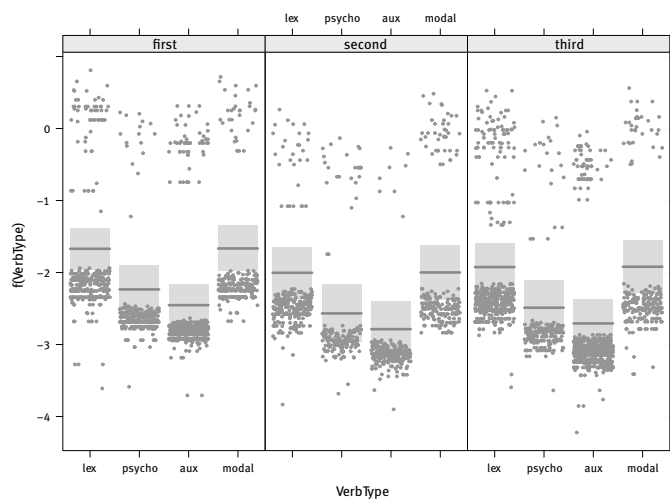


Figure 5.23: Interaction graph verb type by person GSSEC

Combinations not attested at all in other varieties, such as first person zero with modal verb (5.47), or second person zero with primary auxiliaries (5.48), are used regularly in the GSSEC.

- (5.47) C What you_i mean you_i cannot finish?
 A We_i were like just nice you see
 So at the most Ø_i can only finish one and a half
 <GSSEC:056#104:A>
- (5.48) Ø don't want to buy Punjabi suit? <GSSEC:062#1154:C>

A summary of the direction of effects is provided in the effect plots for the full GSSEC model (Figure 5.24).

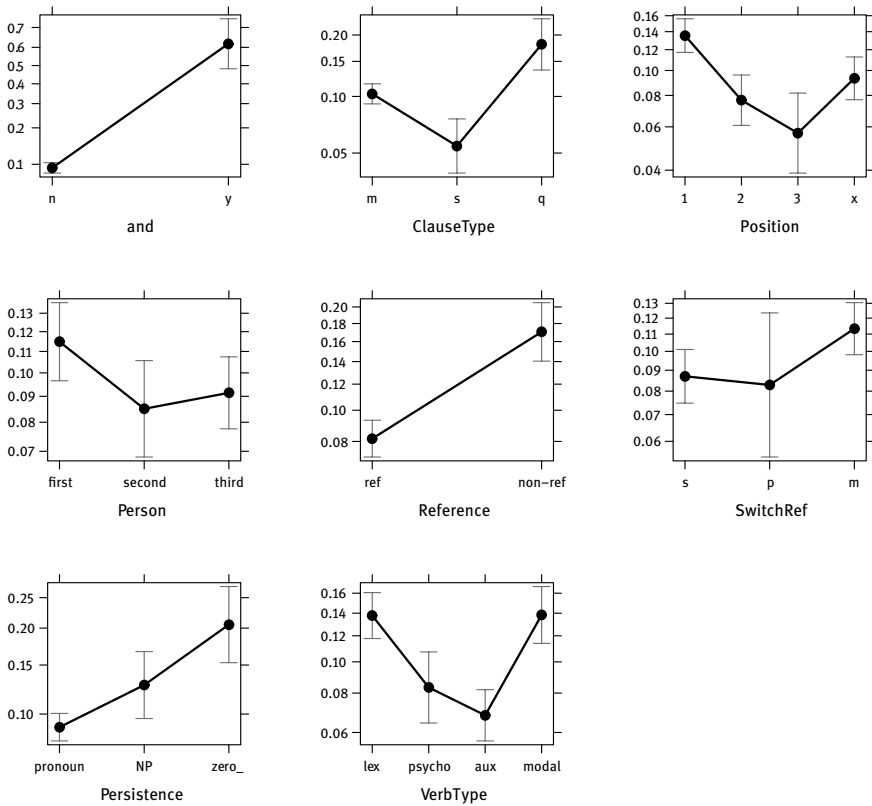


Figure 5.24: Predicted probabilities for factor levels GSSEC

5.5.3 Singapore English: First person contexts

The Singlish data contains 1,186 first person pronouns in variable contexts. Unlike British, Indian and Hong Kong English, Singapore English does not restrict the occurrence of first person null subjects structurally; the omission rate of 11.8% is thus identical to the omission rate of first person pronouns in the full set. Singapore English is the only data set in this investigation that contains null first person pronouns in questions (5.49).

(5.49) I also can, *ah*?

Ø Cannot *ah* three wives? <GSSEC:056#613–614>

In the analysis of the full set of factor groups, coordination, position, persistence and verb phrase influence first person subject realisation significantly, while clause and switch reference do not. The resulting regression model (Table 5.46) is significant in the Model LRT ($p < .01$), but shows very poor discrimination ($C = 0.66$) and explanatory values ($R^2 = .08$).

Table 5.46: Logistic regression model first person GSSEC

	Coefficient	SE	Z	p-Value	
Intercept	-1.6358	0.2015	-8.117	<.0001	***
and: y	2.2544	0.6306	3.575	0.0004	***
Position: 2	-0.7313	0.2818	-2.595	0.0095	**
Position: 3	-0.4930	0.3472	-1.420	0.1556	
Position: >3	-0.4707	0.2411	-1.952	0.0509	.
Persistence: NP	0.2240	0.3092	0.724	0.4689	
Persistence: zero	1.1061	0.2668	4.145	<.0001	***
Verb type: psychological	-1.1415	0.3185	-3.584	0.0003	***
Verb type: auxiliary	-0.2052	0.2255	-0.910	0.3628	
Verb type: modal	-0.1313	0.2734	-0.480	0.6309	

As in the full model, coordination and verb type are the most influential factor groups for first person subjects. In the Singlish data, there is slight evidence of a stronger pronoun favouring effect for psychological verbs in first person contexts compared to the full data set. However, while the model diagnostics reported above suffer from the lower token numbers in all varieties investigated, this blatant loss of validity for Singapore English first person contexts renders further analysis of this subset rather futile.

5.5.4 Topic prominence in Singlish: Within and beyond the clause

As shown above, the distribution of null subjects can be reasonably well explained by a multivariate analysis of the structural factors under investigation here for British, Indian and Hong Kong English. According to the diagnostics measuring the validity of the regression model, this is not the case to the same degree for Singapore English. Given the rather poor explanatory value of the logistic regression for Singapore English null subjects, complementary approaches to the phenomenon are explored. The aim of this section is to pinpoint characteristics of Singlish discourse structure that elude quantitative analysis.

5.5.4.1 Chinese-style topics

Observations by e.g. Ansaldo (2004, 2009), and Bao and Lye (2005) claim a typological shift of Singlish from the English subject-prominent to the substrate topic-prominent language type. As elaborated in section 4.1, this language type is known for the absence of non-referential pronouns and its free usage of null anaphora for both subject and object. The condition on the null realisation of anaphoric elements is described as “discourse recoverability” (Camacho 2013: 125), or referential givenness in context (Li and Thompson 1981: 658). Further diagnostics of topic prominence are double NPs (5.50), a wider range of acceptable topic constituents, such as locatives (5.51) or temporals (5.52), and Chinese-style topics, semantically related to the direct object (5.53) or the whole comment clause, with (5.54) or without overt subject (5.55). These constructions are typically not found in subject-prominent languages, but all of them are attested in the conversations sampled in the GSSEC (see also section 4.1.2, Bao and Aye 2012).

- (5.50) You know then my sister her stupid watch was slow by five minutes
<ICE-SG:015#131:A>
- (5.51) Austria we'll be there for six days <ICE-SG:017#324:C>
- (5.52) Weekends they will open to public <ICE-SG:014#X384:A>
- (5.53) Charles Dickens I read Christmas carols <ICE-SG:090#X402:B>
- (5.54) But durian cake you must have the season *ah* <ICE-SG:006#93:A>
- (5.55) Her Chinese \emptyset_i don't know what she's saying \emptyset_j don't understand \emptyset_i
<ICE-SG:007#187:1:A>

The topic NPs underlined above constitute the typical cases of designating the referent, or setting the frame of the propositional content expressed in the comment. Instead of morphological marking, as for example in Japanese, the topic is marked by word order in Chinese, and Singlish accordingly. While such topic-prominent languages inherently provide this structural initial position for topics in the form of nominal expressions, subject-prominent languages have to “create an adjunct position for it by means of locative marking or lexical specification” (Xu 2015: 402). The typical possessor-possessive relation of double NPs is illustrated by (5.50); a similar metonymic relation is present in (5.53) for the author and his work, and (5.55) for the speaker and her language or pronunciation, while the relation is more indirect in (5.54). Variants following the subject-prominent pattern, e.g. with prepositional phrases as topics, are also part of the Singlish inventory.

English-style topics are attested in Singlish as well, and commonly serve to establish topic continuity (Ward and Birner 2004: 159). These topicalisation or fronting constructions are the result of movement operations, typically involving

the object noun phrase (see also section 2.2.1). The fronted element is anaphorically related to overt or empty pronouns (5.56). According to Winkle (2015), both left dislocation and fronting are not exceptionally frequent in Singapore English. Like Chinese-style topics in (5.55), English-style topics frequently occur without overt subject pronouns (5.56).

- (5.56) 1 B Sometime you_i got the urge to take back organ
 2 A Organ_j Ø_i don't want Ø_j <GSSEC:058#91–92>

5.5.4.2 Topic chains

So far in this section, topic-prominent features of Singapore English have mainly been discussed on the clausal or sentence level. Another distinctive feature of topic-prominent languages are topic chains, which are linked to larger discourse units (see also section 4.1.2). These topic chains are especially common in linear narratives, which are embedded within several conversations in the GSSEC. Narratives typically feature central characters and present their fates and adventures in temporal or logical order, resulting in largely coherent reference in an event-focused unit of discourse (Travis and Lindstrom 2016: 118), so it is no surprise that they favour null pronouns. While this tendency towards zero anaphora in different registers is also observed by Travis and Lindstrom (2016) in a comparison of English conversation vs narratives, and reported by Li (2014) for L2 Chinese, Li et al. (2012) find no significant effect for L1 Chinese. (5.57) presents the start of a topic chain in Singapore English. The narrative featured in the respective conversation presumably starts (shortly) before the recording; the recorded part still follows general principles of discourse organisation.

- (5.57) So what happened was
they had to hand in the project right
 So the deadline was tomorrow or something like that
 <GSSEC:063#1–3>

The extract above sets the goal of the protagonists, as well as the temporal frame, with a formulaic expression as introduction. “They”, later elaborated as “these two guys”, two friends of the narrator, find themselves in a predicament familiar to all students. The rest of the narrative is concerned with the two central characters and their obstacles to handing in their project in time; “they” constitute the discourse topic, and are used as the anchor for a typical narrative topic chain in (5.58).

- (5.58) 1 A They_i took a taxi
 2 B Ok
 3 A Ø_i Told the taxi man to go *lah* one of the roads behind the jungle
 4 Ø_i Asked the taxi man to stop there
 5 so that they_i can walk in *lah* <GSSEC:063#37–41>

This kind of topic chain is commonly used as textbook example, stressing the importance of uninterrupted reference continuity. According to Camacho (2013: 30–31), intervening antecedents block zero anaphora in both discourse- and partial-NSLs, in contrast to canonical-NSLs, where morphological licensing carries the identificational load. In this narrow configuration, a topic chain can be captured well by the factor groups switch reference and persistence, marked as the factor levels “reference maintenance” and “persistence: zero” respectively.

While (5.58) above constitutes a rather straightforward narrative embedded into a conversational setting, dyadic or even triadic conversation is characterised by a much higher amount of referential shift and ambiguity. Still, ongoing zero anaphora reminiscent of topic chains is also found in Singlish conversational data (5.59).

- (5.59) 1 C Then on the second Ø got museum
 2 B I don’t know whether it’s [museum or temples]_i
 3 A Temples uh
 4 B Something I can’t remember
 5 I don’t have such good memory
 6 C Temple *ah*
 7 A Ø_i Cannot be museum
 8 Ø_i Must be temple uh
 9 C Ø Got museum <ICE-SG:011#179–188>

The exchange above presents an ongoing discussion about the deictic reference of “it”. Discussing the destination on day two of a multi-day trip, line 1 presents the proposed referent (“museum”) in the subjectless existential *got*-construction. Speaker B questions the propositional truth of this statement, offering the alternative referent “temples”. Even though the antecedent is not clear concerning its actual referent, it is kept consistent, and zero anaphora is still possible in lines 7 and 8 by speaker A, who asserts “temple”, whereas speaker C insists on “museum” to be the correct referent (the question is never resolved conclusively in the transcribed conversation).

While referential continuity is a prime enabler for zero anaphora in narrative topic chains as in (5.58), in natural conversation reference is usually shifting

quickly and continuously, and increasingly so with higher interactivity between speech participants. The presence of intervening antecedents pushes the boundaries of the rigidly delimited analytical categories defined for the quantitative analysis, especially those trying to account for linguistic context in the form of referential continuity of denotations, and structural persistence of formal means of expression.

5.5.4.3 Discourse orientation and ambiguity

Besides the features discussed above, topic prominence, or pragmatic orientation of a language, entails a less rigid sentence structure, and demands more interpretative work from the hearer (e.g. Huang 1984). Not only are Chinese-style topics a more open class semantically, they can also stand in an ambiguous relation to the verb phrase, which can only be resolved from the discourse context (5.60).

- (5.60) a. *Zhè jī bù néng chī le.*
 DEM chicken cannot table eat
 ‘This chicken cannot eat anymore (because it is sick)’
- b. *Zhè jī Ø bù néng chī Ø le*
 DEM chicken cannot eat LE
 ‘(one / you) cannot eat this chicken anymore’

(Loar 2011: 9)

The interpretation in variant a. requires a conflation of topic and comment with subject and predicate of the clause, i.e. the “chicken” in this case constitutes both the topic and the agentive subject of the clause. Variant b., on the other hand, presumes an empty pronoun acting as the implied generic subject of the comment clause. This clause provides information about the English-style topic “chicken”, which is anaphorically related to an empty pronominal in object position (see also 5.56). The corresponding structure is also found in Singlish (5.61, curiously enough with almost identical content as 5.60, see also Gil 2001 on Riau Indonesian).

- (5.61) Two metre crab can eat one.
 <ICE-SG:085#391:A>
 a. This two metre crab can eat (some)one.
 b. (One / you) can eat this two metre crab / This two metre crab is edible.

Again, variant a. constitutes an interpretation based on canonical English SVO structure, with an initial NP as agentive subject, and the indefinite pronoun *one* as direct object. However, *one* represents another instance of grammatical replication in Singlish, acting as a multifunctional marker of nominalisation (5.62), singulative, i.e. highlighting a unique property of the subject (5.63), as emphasiser (5.64), and relative particle (5.65, Bao 2009, Wee and Ansaldo 2004: 68–71; see also section 4.3.3).

- (5.62) Cannot finish *one*.
(Speaker tells addressee that the packet of peanuts he bought recently contains so many that he has difficulty finishing them)
(Wong 2005: 257)
- (5.63) You always late *one*!
(Wee and Ansaldo 2004: 70)
- (5.64) You have to be very careful *one*. Otherwise, you will lose *one*.
(Wong 2005: 249)
- (5.65) The boy pinch my mother *one* very naughty.
'The boy who pinched my mother is very naughty'
(Alsagoff and Ho 1998: 135)

The Singlish *one*-relative clause is a mixed structure, calquing the functions of Mandarin particle *de* on the English lexeme *one*, but following the English syntactic pattern of post-nominal modification (Alsagoff and Ho 1998: 132–133). The interpretation in variant b. of (5.61) is strengthened by the common co-occurrence of the lexeme *CAN* with generic null subjects, such as (5.66). Instances of *can* and *cannot* without overt subjects contribute significantly to null subjects in Singapore English; other modal auxiliaries are not used to the same degree (see also Table 5.5).

- (5.66) The shell Ø can eat
'The shell can be eaten'
(Wong 2014: 141)

Wong describes *can*, and its negated form *cannot* as “ubiquitous” in Colloquial Singapore English (Wong 2014: 139). Beyond overlapping functions with the English source expression, Singlish has further functions for its “cultural keyword”: *can* is commonly used in the place of *could*, presumably evoking “the impression of ‘interest and involvement’”, compared to the more hypothetical and “detached” Anglo English version with *could* (Wong 2014: 144–145). According to Wong’s analysis, *can* is especially common in interrogatives and interrogative-directives.

The sociocultural implication seems to lie in the disregard for the addressee's volition, compared to their ability to fulfil the request (Wong 2014: 171). Singlish *can* / *cannot* thus represents a higher degree of imposition in a request. (5.67) illustrates this pragmatic aspect of *cannot*, by contrasting the invariant tag *can or not* with the more literal use of *cannot* as answer.

- (5.67) 1 B But then I mean it's very difficult to to reject people whoa.
 2 I never say anything <unclear> people call you already.
 3 Hello Ø go to your house can or not?
 4 Ø Cannot.
 5 You say Ø cannot?
 6 A No *lah*.
 7 You say I'm not free. <ICE-SG:007#218–223>

Speaker B talks about the difficulty “to reject people”. Line 3 cites an inquiry or request by “people”; it is formulated without overt subject, and adds the invariant tag *can or not* to inquire about the possibility of a visit at B's place. However, the literal answer *cannot*, mirroring the structure of the tag, is deemed unacceptable, although it is obviously the preferred statement by B. Given the regular usage of this expression as a complete answer in Singlish, this unacceptability is not based on syntactic grounds, but rather on pragmatic reasons. Reassessing its acceptability by asking A, speaker A agrees that “cannot” is not a suitable response, suggesting the more indirect statement “I'm not free” as an alternative. This formulation shifts the emphasis away from the non-permission, onto the inability of the speaker to fulfil the request, an interpretation that is in line with Wong's characterisation of Singlish *can*-requests described above.

Although *can* and *cannot* are the most prominent of these expressions, (Wong 2005: 246) states more generally that “Singapore English speakers often use modals as abbreviations for sentences”, illustrated by (5.68). The context provided is “[t]he addressee is worried that the speaker may get hungry at night, but speaker tries to convince him otherwise. [...] *won't* may be expanded into ‘It won't happen’” (Wong 2005: 246).

- (5.68) Won't one la .
 bù *hùi* *de* *lǎ*
 not will DE LA

(Wong 2005: 246)

This expression is one of many examples of a direct calque of Mandarin Chinese structure in Singlish.

Predicative adjectives, or zero copula, is another one of the features assumed to have transferred to Singlish (Ansaldò 2004: 135), and indeed, copula and auxiliary BE omission is frequently encountered in Singlish conversations (5.69, see also section 5.1.1).

(5.69) I Ø also scared <GSSEC:031#123:B>

Given the blurred boundaries between the categories of predicative adjectives and stative verbs in Chinese (Ansaldò 2009: 141), it is likely that these contexts are included in the quantitative analysis of Mandarin Chinese; while this is not explicitly addressed in Jia and Bayley (2002), it is evident in (5.70) from Li et al. (2012).

(5.70) *Ránhòu nǐ shēngbìng le Ø dōu. tèbié dānxīn*
and you sick LE Ø all very worried

‘And when you were sick we were all very worried.’

(Li et al. 2012: 99)

However, Chinese does use overt copula for emphasis (Li and Thompson 1981: 149–155), its lack is thus not fully categorical. Copula omission is variable in Singapore English as well; in fact, the majority of possible contexts occur with overt copula in Singlish (Ho and Platt 1993).

5.5.4.4 Repair

Anaphoric reference involves a tension between economy of production and clarity, i.e. the most explicit linguistic reference (Huang 1994: 216). Felicitous use of (zero) pronouns is thus dependent on the lack of referential ambiguity. Unresolved ambiguity in conversation usually results in repair strategies, either speaker- or self-initiated, or alternatively listener- or other-initiated, and conducted by either speaker or listener (Wales 1996: 23). Self-initiated self-repair is the most common type, typically consisting of an explicit mention of the referent in the form of a full noun phrase (5.71, *he* is the *reparandum*-pronoun, the *repairs*-NP is marked in bold, see also Huang 1994: 216–218).

(5.71) So although he realised **Finn** realises his dream of being an artist
<ICE-SG:090#196:A>

The bold part in (5.72) below illustrates other-initiated repair, requested (“you mean”) and conducted by the listener (contrastive pronoun *he* underlined, line 11).

- (5.72) 1 A Everyone outside keeps thinking
 2 that I am always interested in all their sort of work
 3 as long as there are statistics involved
 4 Ø Got one
 5 that wanted me to collaborate with him on <unclear> cancer
 6 Ø Wants me to go into the cancer
 7 pull out all the cases uh that’s <unclear> cancer
 8 and and relate it to hormonal level
 9 B In this hospital
 10 A Ya ya ya he’ll
 11 B **You mean he wants you to do the pulling out**
 12 A Ah yes yes yes
 13 B Wow good you said no
 14 A Ya I said no <ICE-SG:033#297–305>

Speaker A complains about being too busy professionally, singling out one colleague among several (line 1 “everyone outside”) who keep adding to A’s workload. This individual is introduced via the existential *got*-construction in line 4, and elaborated by the following relative clause in line 5, a common pattern of referent introduction. This person is the obvious referent for the zero anaphora in line 6; its identification is aided by third person marking of the verb phrase. However, the interpretation of the canonical null subject of the infinitive in line 7 seems more problematic, evoking the explicit request for clarification of the agent of the verb “pull” by speaker B in line 11. Similar to (5.67), the source of the unacceptability is based on pragmatic rather than syntactic reasons; speaker B primarily expresses their incredulity about the request.

These types of ambiguous anaphora occur in any naturalistic conversation, and obviously in Standard English as well (Wales 1996: 23). Although the potential for misunderstanding seems higher when zero anaphora is included as a regular option, such examples featuring explicit repair mechanisms are rare in the GSSEC. Moreover, like in (5.72) above, zero anaphors repeatedly appear in clusters. This is also illustrated by (5.73) below, where, besides multiple null subjects, several instances of null objects (line 5, 10, 11) and zero copula (line 3) occur in close neighbourhood.

The conversation is concerned with timetables for the coming term, and how to get access to them. Lines 9, 10 and 11 constitute a cluster of three consecutive

null subjects. These are most likely coreferential, referring to speaker B, although line 10 and 11 also allow for a potential generic reading, discussing the location of the timetables in more general terms.

- (5.73) 1 B Uhm no I_i have a rough idea
 2 but I_i must see whether they clash or not
 3 A Timetable Ø out you know
 4 B Is it
 5 Who said Ø
 6 A That's how [name] knows his clash *what*
 7 B I_i couldn't access it from home
 8 A You_i can go and get [a copy]_i
 9 B Ø_i Can go later *lah*
 10 A Ø_i Take Ø_j from where
 11 B AS Seven Ø_i can get Ø_j <ICE-SG:085#143–151>

5.5.4.5 Code-switching?

The frequent occurrence of such clusters of null subjects, commonly in combination with other features typical for the colloquial variant Singlish such as topicalisation, *or not* question tags, null objects, *wh*-in-situ questions, lack of morphological marking, copula omission, and discourse particles, raises the question whether this is an instance of structural priming, as identified for English (see chapters 2 and 3), or whether they manifest instances of stylistic shifts, be it code-switching, or indexical usage of linguistic resources of either substrate or superstrate provenance (Leimgruber 2013: 113).

Another factor is the relatively high amount of idiosyncrasy for the Singaporean speakers. While there are only two speakers with invariant overt subject realisation (see section 5.2.1 above), compared to the other data sets the Singaporean data exhibits higher internal variability of amount of null subjects, not only per conversation (see Figure 5.5 above), but also per speaker (Figure 5.25).

This is in line with Leimgruber's (2013: 126) observation on the importance of analysing idiolectal speech to grasp the differences in situational use of linguistic resources in dynamic, emerging varieties like Singapore English. On the other hand, compared to the L2 Asian varieties, British English also exhibits a higher amount of idiosyncratic behaviour in the form of numerous outlier speakers, deviating significantly from the more typical speakers regarding their subject omission rate. Given the lack of proficiency issues as a possible explanation for the internal variation in Singapore English, this parallel can be seen as a further

indicator of L1 status, leading to higher internal diversification, compared to more standard-oriented L2 varieties (Schneider 2007: 18).

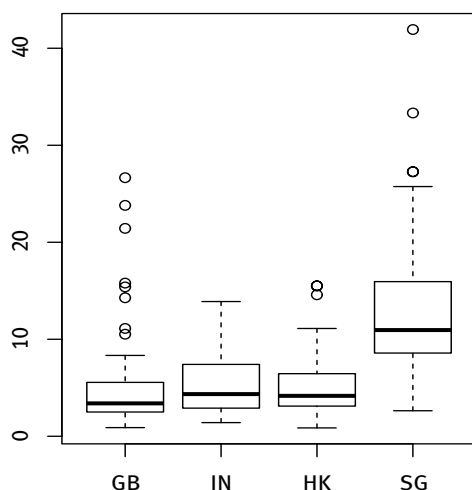


Figure 5.25: Comparison percentage null subjects by speaker

Seemingly random variation between standard and indigenised variants is also described by Lim and Ansaldo (2016: 53–55) for classroom speech of both students and teachers. Multilingual Chinese-English students “[show] a flow between Standard Singaporean English, Singlish and Mandarin, with a lack of neat, discrete boundaries”. The determinant of this variation is found in the appropriateness “in the interactional moment”, conditioned by the different learning tasks (Lim and Ansaldo 2016: 55). This interpretation is strengthened by observations on teacher speech, which show the teacher using standard features for utterances concerned with lesson organisation, while focus on teaching content triggers a higher amount of localised features (see also Alsagoff 2010). This dependence of subject realisation on the discourse situation mirrors findings on Mandarin Chinese null subjects, discussed in section 4.1.3.

5.5.5 Short summary: Null subjects in Singapore English

The analysis of the Singapore English data has shown that the extra-linguistic factor ethnicity does not significantly influence subject realisation, confirming observations by Deterding and Poedjosoedarmo (1996) and Lim (2001), who iden-

tify ethnic differences predominantly for lexical and phonological features, especially intonation, but decidedly less for grammatical features. Speakers largely belong to the younger generation of Singlish L1 speakers (Lim and Foley 2004: 11), and are thus not assumed to exhibit differences in proficiency. A comparison with older speech data, or with older speakers of Singlish, representing the L2 stage of English in Singapore, is desirable, especially regarding the comparison with the L2 varieties Hong Kong and Indian English.

The regression model is not a good fit for the Singlish data. For both the full data set as well as for the first person subset, the model evaluation yields poor diagnostic values for model discrimination, and the amount of variation explained by the model. Within their limited predictive power, the strongest effects are measured for the factor groups coordination and verb type, followed by the substrate-inspired factor group specific reference. Switch reference and person are least influential. There is no clear dominance of either subgroup of factors for Singapore English null subjects, although several cases of replica grammaticalisation are attested in the Singlish data (see Heine and Kuteva 2005: 93), e.g. existential *got*, bare conditionals, *X or not* questions, and *one* as nominaliser and relative particle.

Given the limited explanatory value of the regression model, section 5.5.4 draws a larger picture of the widespread structural influence of topic-prominent languages on Singlish, concluding that not all aspects of topic prominence are quantifiable through the analytical categories tailored to subject-prominent European languages. Finer-grained situational analysis, employed in indexical models like Alsagoff's (2010) cultural orientation model and Leimgruber's (2009, 2013) analysis of stylistic variation in Singapore English can help to account for the "seemingly random variation" (Leimgruber 2013: 113) found in current Singapore English with regard to the frequency of null subjects, especially between individual speakers. Still, the comparative typological approach of the present study helps to identify and circumscribe the origins and structural contexts enabling the occurrence of null subjects in the first place.

5.6 Summary and Discussion: Null subjects in Asian Englishes

For the sake of comparability, the analysis of null subjects in the Asian varieties necessarily adheres to the guidelines established for British English in chapter 3. Following the comparative variationist method, the blueprint regression model fitted for Standard English is projected onto the Asian Englishes to gain insight into the currency of variable rules for subject omission in the respective varieties. The comparison of British English and the Asian Englishes is conducted in

chapter 6; this discussion provides an evaluation of the results presented in the preceding sections on the individual Asian varieties, including a comparison of the group of Asian Englishes (sections 5.6.1 and 5.6.2), and a contrastive look at linguistic constraints established for the topic-prominent, or radical NSL Chinese (section 5.6.3).

5.6.1 Summary: Asian English null subjects

The tenet of variationist analysis is that rule-governed differences between grammatical systems can be observed through the quantitative variation of surface forms. A first insight is granted by divergence in the envelope of variation of the linguistic variable in individual varieties. Circumscribing the variable context is a crucial step in each variationist analysis. It requires a careful balance of excluding invariant, and therefore irrelevant contexts, and including variant, albeit less central contexts. It is these niches that constitute structural innovations in contact varieties, and thus might prove particularly insightful, especially in comparative analyses. The relatively broad variable context defined for the analysis of British English in section 3.1 is justified by the observations in section 5.1: generally speaking, null subjects in Asian Englishes are less rigidly constrained by structural contexts than their Standard English counterparts (see also Torres Cacoullos and Travis 2014 on American English). As expected, Singapore English is the most conspicuous maverick structurally, exhibiting numerous constructions that represent direct calques from the local substrates, including their preference for null subjects (see also section 5.5). However, all three Asian varieties show a number of productive categories for null subjects that are more peripheral for Standard English, especially concerning the category clause type, such as Hong Kong English relative clauses. Still, the divergent patterns in each of the three varieties concerning the wider envelope of variation cannot be unified into targeted influence of substrate configuration on specific contexts, but rather emerge as variety-specific constructions, gradually weakening the near-categorical superstrate constraints.

The descriptive account of the corpus findings in section 5.2 shows that the patterns of null subjects within the corpora differ with respect to the language-external factors *conversation*, *individual speakers*, and regarding *collexemes*. Across conversations, Indian English is more focussed than Hong Kong English and especially the Singlish data, which both show more heterogeneous behaviour. However, none of these factors significantly improves the multivariate models when added as random intercepts. On the other hand, null subjects are more widely attested in the Hong Kong and Singaporean speech communities, as indi-

cated by the lower number of speakers with categorical overt subject realisation (see Figure 5.4) – null subjects seem to be a stable feature of the two Southeast Asian varieties. In terms of overall subject deletion rates, the Singlish data is closer in frequency to its substrates than the other two Asian varieties; Indian and Hong Kong English have almost identical deletion rates that are relatively close to those found in the British English superstrate.

A common result of previous studies is the preponderance of structural over extra-linguistic factors influencing subject realisation. Accordingly, sociolinguistic factors like age (tested for ICE-IN), gender (tested for ICE-IN and ICE-HK), and (co-)L1 (tested for ICE-IN and the GSSEC) do not show a statistically significant influence in either variety. In dynamic situations of language contact, this provides further insights on the status of a structural feature in an emerging grammatical system: it seems null subjects are neither a change in progress, nor do they present an interference feature for learners of specific L1s. Age as a sociolinguistic factor is potentially more interesting for Singlish and Hong Kong than for Indian English – the linguistic situation in India is characterised as relatively stable (see section 4.3.1). In contrast, the sociolinguistic developments in Hong Kong (increasing proficiency in English over the last decades, but possible changes due to current political circumstances, see section 4.3.2) and Singapore (a growing tendency towards English as L1 and home language, see section 4.3.3) are more dynamic. Given the development of Singapore English toward endonormative stabilisation (Schneider's Phase 4), ethnic differences were not expected; grammatical homogeneity among the different ethnic groups is well described, and moreover, the different substrates present in Singapore both historically and presently converge with regard to the syntactic configuration topic prominence (see section 4.3.2). However, ethnically Malay speakers have the lowest average deletion rate, and ethnic differences in Singapore English might be worth further study beyond the limited set of speakers investigated here, considering the varying degrees of the use of English as a home language in the different ethnic groups.

Differences between the three Asian varieties are partly based on speech style. In the case of World Englishes, this can involve differences in the depth and range of linguistic repertoires available (see e.g. Schneider 2007: 82–83). Considering the conversation topics, as well as the rapport between speakers, and the usage of informality markers like discourse particles and non-standard vocabulary, such as cuss words and indigenised expressions, the Singlish data is more colloquial in nature than the Indian and Hong Kong English conversations: the data from ICE-IN features several semi-formal, interview-like interactions, while conversations in ICE-HK commonly include native speakers of Standard English, or speakers from outside Hong Kong. It is unlikely that the kind of

informal data representing Singlish in the present study is even available for a large share of the Indian and Hong Kong English speech communities due to the more restricted functions in the respective societies, and the lower amount of L1 or home language users (see also Percillier 2016: 192 on determinants of stylistic choices in different types of contact varieties of English, and Hansen 2018: 98–99 on the demographic composition of the different ICE components influencing the speech style represented). The higher degree of internal diversification characteristic for more mature contact varieties is also a possible explanation for the higher amount of individual speaker variation found in the Singaporean data, which is closer to the wider distribution found in Standard English than the more homogeneous behaviour of proficient L2 speakers in Hong Kong and India.

Collexemes and high-frequency verb tokens show varying influence on subject realisation in the different varieties. Considering both verb lemmas and verb forms, only forms of GO and LOOK emerge as shared collocates of null subjects in all three varieties. This tendency is stronger for verbs showing significant repulsion of null subjects, which are dominated by (contracted) forms of BE and HAVE for all varieties. Additionally, Hong Kong and Singapore English share *think* as a high-frequency verb form highly unlikely to occur with null subjects. Different verb forms emerge as high attractors for null subjects: *means* for ICE-IN, *is* for ICE-HK, and *got*, *depends* and *can/cannot* for the GSSEC. Collocations are most influential in the Singaporean data, i.e. they have consistently high collocational diagnostics, and include a larger group of verbs. Interestingly, the same is true for verb forms significantly repelling null subjects; Singapore English exhibits the clearest values for repulsion, and includes the largest number of individual verb forms, including several forms of DO. This is in line with findings on *collostructional nativisation* (Mukherjee and Gries 2009): a higher degree of sociolinguistic nativisation is reflected in greater deviance of collocational patterns from the superstrate. Among the Asian Englishes investigated here, Singlish is clearly at a more advanced stage concerning collexemes of null subjects than the L2 varieties Indian and Hong Kong English. Influential lexical attractors of one variety are generally not attested for the other two; apparently these collocational preferences for null subjects are highly variety specific, while the pronoun preserving effect of the high frequency primary auxiliaries BE and HAVE is valid more generally.

Concerning the structural factors identified as influential in chapter 3, different subsets of factors are relevant for the different varieties. An overview of the strength of factor groups for the Asian varieties, based on the Wald statistics of the respective regression models, and the range of the factor group yielded by the Rbrul centred factor weights, is provided in Table 5.47. The table shows statistical

significance¹², and the range of the factor groups (subtracting the lowest factor weight from the highest factor weight within a factor group, see Tagliamonte 2006: 242). Non-significant factor groups are provided in square brackets, the top three factor groups for each variety are marked in bold.

Table 5.47: Factor group significance and range: Asian Englishes

	Factor groups	IN	range	HK	range	SG	range
English-specific	Coordination	*	57	*	55	*	59
	Clause type	*	54	*	33	*	32
	Position	—	[19]	—	[18]	*	25
Contact	Person	*	26	*	32	—	[09]
	Specific reference	—	[03]	—	[11]	*	21
Universal	Switch reference	*	51	*	25	.	[10]
	Persistence	*	63	*	38	*	23
	Verb type	*	62	*	36	*	19

Coreferential coordination is the only factor group highly significant for all three varieties. Combining the favouring effect of referential continuity on cognitive accessibility of the referent, and the situational unity, at least for hendiadic coordinations, with *and* as an explicit marker of subject continuity, it is no surprise that the referent is highly predictable in this context, and therefore frequently requires less explicit linguistic coding. For both Singapore and Hong Kong English, this factor group is in fact the most influential one, based on the range of the factor group. Clause type is also a significant predictor in all varieties, with declarative main clauses as the most likely context for null subjects; this tendency is strongest in Indian English, where subordinate clauses are highly unfavourable towards null subjects. Among the Asian Englishes, the position constraint is most influential in Singapore English. While position 2 and 3 show a slight disfavoured effect for subject omission in all varieties, this tendency is not statistically significant in either Hong Kong or Indian English. Both varieties recurrently permit introductory phrases like *and* or *and then* with both overt and null subjects. The degree of weakening of this strong superstrate constraint is remarkable, especially given the fact that it is not attested to the same extent in the most indigenised variety Singapore English.

12 ' ' not significant, '.' barely significant with $p < .1$, '*' significant with $p < .05$

The contact factor groups person and reference provide a mixed picture. Person crucially determines subject realisation in Indian and Hong Kong English, and the patterns identified are parallel: both varieties favour third person null pronouns, while first person omission is least likely. Second person shows intermediate values, although it is less common in ICE-IN than in ICE-HK. In Singapore English on the other hand, first person pronouns are most likely to be omitted, while second and third person are both less common, and relatively similar in their predicted probabilities of omission. There is no immediately striking explanation for this, but it is remarkable that, again, Hong Kong and Indian English show similar constraints, while Singapore and British English behave differently from the L2 varieties, but similar to each other (Figure 5.26, compare Figure 3.15 for predicted probabilities in ICE-GB).

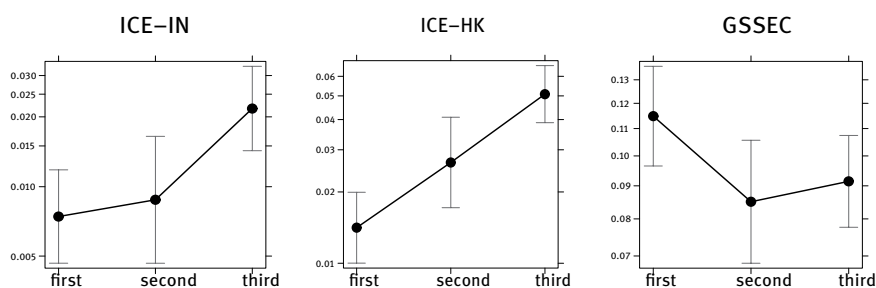


Figure 5.26: Factor group person in the Asian Englishes

Specific reference is surprisingly ineffective as a predictor for null subjects in Asian Englishes, given its categorical absence for expletive use, and the reported rarity in generic contexts in the topic-prominent substrate languages (see section 4.1.2). Still, there is no measurable effect of referential status for Indian and Hong Kong English, and the favouring effect of non-specific reference for null subjects in Singapore English is limited. The widespread presence of overt expletives is described as a relatively sudden development in Mesthrie's comparison of SAIE basilect vs acrolect: while the basilect exhibits 57% zero expletives, this feature is categorically eradicated from the acrolect (Mesthrie 1992: 170, see also section 4.2.4), a shift that is explained by the switch between the discrete parameter settings "pro-drop" for the basilect, and "non-pro-drop", requiring overt expletives, for the acrolect. This contrast is also reported by Bhatt (2004) for basilectal vs acrolectal Indian English. Interestingly, this discrepancy is parallel to observations made on L1 development of English-speaking children (Valian 1991). Appar-

ently, the assumed parametric shift affects the categorically empty category of expletive subject pronouns more effectively than anaphoric referential pronouns, which are more variable in their realisation in the substrate languages.

The universal factors persistence and verb type are highly influential in all three Asian Englishes; in Indian English, they bear the greatest statistical weight, for Hong Kong English, they follow right after coordination. With regard to persistence effects, the relative rarity of overt coordination (Figure 5.6 above) is possibly tipping the scales in favour of the factor group persistence in these varieties. ICE-India prominently features several extracts of speakers listing their activities within a given timeframe, e.g. describing their morning routine, in consecutive clauses, without overt coordinators (5.74).

- (5.74) I don't know how I'm going to manage it
 Everyone just rush up <,>
 Ø prepare everything <,>
 then Ø take our lunch <,>
 then <,> Ø prepare lunch for our children <,>
 <ICE-IN:030#149–154:A>

These lists resemble topic chains found regularly in topic-prominent languages (although they are not exclusive to this language type). The effect of persistence is not as decisive for Hong Kong, and even less so for Singapore English. Apparently, in these varieties null subjects are possible with intervening referents to a higher degree, diminishing the favouring effect of immediately preceding zero (see also section 5.5.4 on discourse-level patterns of Singlish). Regarding verb type, Indian English shows a relatively extreme aversion to null subjects with both primary and modal auxiliaries. While this tendency is also found in the other varieties to a degree, the effect is much stronger for Indian English. Again, verb type is not as central for Hong Kong, and even less influential in Singapore English. Both of these universal factor groups are more important in Indian and Hong Kong than in Singapore English; the same is true to a lesser degree for the third universal factor, switch reference, which shows the highest factor weight in Indian, followed by Hong Kong, and is barely touching statistical significance in Singapore English.

Both ICE-IN and ICE-HK can be well described by the present model, as is evident from the diagnostic values of model significance provided above (sections 5.3.2, 5.4.2). The L2 varieties also exhibit the same set of significant factor groups; the English-specific factor position and the contact factor specific reference are not significant in either. This is different for Singlish, where position and specific reference do contribute significantly, while the contact factor person

and the universal factor switch reference are rendered insignificant. Moreover, variation in subject pronoun realisation is described less accurately by the multivariate analysis of the Singaporean data, raising the question which factors suitable for operationalisation in a quantitative framework can possibly capture the variation more insightfully in this variety. Concerning the diagnostic measure of factor group significance, it is thus not Hong Kong and Singapore English that pattern, based on shared or at least more closely related substrates, but rather Hong Kong and Indian English, which show more parallels regarding their nativisation status, and the frequency and domains of usage in the respective societies.

5.6.2 Variety specific contexts and puzzles

As shown in the comparison above, the Asian varieties show divergent behaviour regarding a number of structural constraints, especially person (Figure 5.26) and verb type. The contact factor person is most relevant for Hong Kong English; especially the lack of first person zero subjects is conspicuous and singular to this variety (differences within this specific context are further discussed in section 6.1.2). The clear contrast between different persons in Hong Kong English, and the high likelihood of third person null pronouns, resembles the split system of partial NSLs, as described in sections 2.1.3 and 4.2.4. Given the status of third person singular verb phrases as the only context potentially identified by inflectional morphemes, it is worth considering morphological marking as an enabling factor in future studies on Hong Kong English; however, Tamaredo and Fanego (2016) fail to find significant effects in ICE-IN and ICE-SG. The divergent behaviour of Singlish regarding the verb type modal auxiliary is, at least partly, due to direct substrate calques, especially with *can/cannot* (see also section 5.5.4); concerning the remaining verb types Singapore English behaves similarly to the other two Asian Englishes, which show parallel rankings of the factor levels within both factor groups person and verb type. Besides person and verb type, the strongest contrast between the Asian varieties is visible for the factor group clause type (Figure 5.27).

Clause type is identified as a significant factor group for all varieties in this study. Singapore English is the only variety to exhibit a robust tendency towards higher likelihood for subject omission outside declarative main clauses, i.e. in questions. While the results for Singapore English questions are statistically valid, in Indian English question is not significant as a factor level, qualifying the slight favouring effect of this context on null subjects (see Table 5.13); the relatively high likelihood is based on the overall low occurrence of questions in the Indian data (see Figure 5.1).

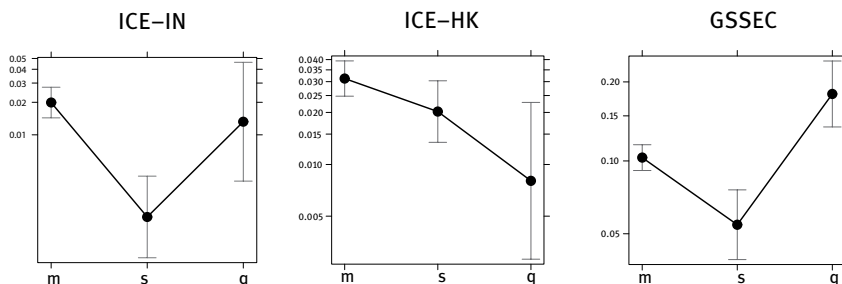


Figure 5.27: Factor group clause in the Asian Englishes

In Singapore English, on the other hand, substrate calques like *X or not* or *wh*-in-situ questions are productive variants of question formation, and frequently entail the use of null pronouns. Hong Kong English clearly disfavours null subjects in questions, but the local preference for null relative pronouns in subject position waters down the superstrate constraint on clause types. According to Gisborne (2000), null subject relativisers are so common in Hong Kong English that local speakers hesitate to even identify them as errors. This lack of stigmatisation is a possible explanation for the higher frequency of null relative pronouns, especially regarding the comparably low frequency in the Singaporean data. Furthermore, null relativisers are described as a learner phenomenon of Chinese L1 speakers e.g. by Hung (2012), and thus more likely to be encountered in a sociolinguistic setting with ongoing L2 acquisition. Despite these divergent patterns in the Asian Englishes, overall, clause type as a factor group is still more justified for the Asian varieties, Singapore English especially, than for British English, where the predicted language specific constraint to declarative main clauses is indeed stronger. It is still remarkable how differently substrate structures manifest in different contact varieties, especially when the substrates show parallel or even identical patterns, as is the case for many structural aspects for varieties of Chinese.

5.6.3 Substrate comparison: How Asian are Asian Englishes?

The analysis of structural constraints in both superstrate and substrate languages is necessary for the balanced evaluation of contact effects. Mandarin Chinese is the only substrate, and in fact the only topic-prominent language, for which quantitative analyses similar to the one conducted here are available. However, a

more detailed look shows how difficult it is to compare studies of eclectic provenance: they commonly follow different principles in the definition of the variable context, and contribute mismatched, or even conflicting analytical categories. Moreover, unfortunately no diagnostics of model fit are made available for the studies on Chinese null subjects by Jia and Bayley (2002) and Li et al. (2012). Since this aspect poses a large question mark behind the results on Singapore English achieved in the present study, the variety structurally closest to Chinese, this constitutes a major loss in comparability.

Meyerhoff's (2009) comparative variationist approach calls for a juxtaposition of structural constraints in contact varieties with those of their substrate. She proposes three diagnostic conditions, which describe the extent of matching constraints in two or more languages being compared, and give insight into the kind and quality of transfer. *Weak transfer*, or replication, is indicated by the same significant factor groups in all models, *strong transfer* by the same order of factor groups regarding their statistical significance, and *calquing*, the highest degree of transfer, by the same order of factors within factor groups (Meyerhoff 2009: 304, see also Figure 4.1).

However, this approach presupposes the same analytical categories for all languages and varieties included in the comparison. In the Chinese studies, the English-specific factor groups coordination and position, and the supposed universal factor groups persistence and verb type, which are highly significant for the Asian Englishes, are not analysed. Furthermore, Jia and Bayley (2002) consider human subjects only, hence the factor group reference, or specificity, is only available for Li et al. (2012). The following comparison is thus limited to the comparison of the factor weights, and the order of factor levels available.

Table 5.48 presents a matrix of similarity of Chinese and the Asian Englishes, based on the available results of multivariate analyses, providing significance and range of the factor groups, as well as the order of factor levels. Non-significant factor groups are provided in square brackets for the respective varieties.

Clause, or sentence type, and person and number are the most influential factors for Chinese null subjects. A remarkable parallel is the favouring effect of questions on null subjects for both Singapore English and Chinese; for Li et al.'s data, at least there is no marked difference between statements and questions, a clear contrast to English. Clause type, specifically questions, thus constitute a likely instance of Meyerhoff's highest degree of substrate transfer, calquing, in Singapore English.

Table 5.48: Factor range and ranking of factor levels Asian Englishes - Chinese

		IN	HK	SG	Jia and Bayley 2002	Li et al. 2012
English-specific	Tokens overall	3,636	3,979	3,939	1,400	8,507
	% zero overall	5%	5%	12%	47%	53%
	Clause type	***	*	**	***	**
	range	54	33	32	30	26
		main	main	q	imperative	imperative
		q	sub	main	q	main
Contact		sub	q	sub	main (statement)	& q
	Person	***	***	–	***	***
	range	26	32	[09]	43	36
		3rd	3rd	1st	‘context marked’ 2sg & 2pl	2pl & 3pl
		2nd	2nd	3rd	3pl	3sg [-an]
		1st	1st	2nd	2pl	1sg & pl, 3sg [+an]
					1pl	2sg
					2sg & 3sg	
					1 sg	
	Specific reference	–	–	**	NA	*
Universal	range	[03]	[11]	21		06
		ref	non-ref	non-ref		non-spec.
		non-ref	ref	ref		specific
	Switch reference	***	***	.	**	***
	range	51	25	[10]	13	43
		switch	m	m	m	m
		m	partial	switch		m
		partial	switch	partial	switch & partial	(‘same underlying’) switch partial

It is worth noting that for the factor group clause type, both Chinese studies include imperatives, considerably broadening the scope of possible subject omission and accounting in part for the high rates of subject omission (47% and 53%, respectively). Obviously, imperatives constitute a context highly favourable to subject omission. Imperatives with overt subjects amount to roughly 30% in both studies; apparently this option is more prevalent than e.g. in English. Evi-

dently, including imperative contexts for the Asian Englishes would raise their null subject rates considerably, narrowing the apparent gap of null subject rates to Chinese. Imperatives are restricted to second person contexts, but it is not clear how this evident interaction between the factor groups clause and person is handled in the Chinese studies. The influence of this syntactic context on the manifestation for different persons undermines the comparability of the factor person between Jia and Bayley (2002), Li et al. (2012), and the present study.

However, despite the inclusion of imperatives, second person singular omission in Chinese is comparatively rare, albeit heavily dependent on discourse context, i.e. the situationally dependent predictability of the omitted referent: in teacher speech, second person plural is the more likely variant, compared to telephone conversations, where second person singular is the most likely addressee. Such situational differences are not as relevant for the Asian Englishes data analysed here; still, they share a common tendency towards higher rates of overt subjects for second person contexts. The relatively low amount of second person zero in Li et al. is all the more remarkable considering they also include generic, or non-specific, pronouns in this context. Furthermore, given the inclusion of number distinctions for the Chinese study, and the apparent post-hoc collapsing of categories within this factor group into non-intuitive subgroups, further generalisations are not possible based on the data available. One tendency, however, deserves specific mention: first person contexts, the least likely context for pronoun omission in Indian and Hong Kong English, also exhibit a pronoun-preserving tendency in Chinese; the differences between the Asian varieties in the constraint hierarchy for this factor group are thus a possible indicator of substrate convergence.

The most surprising result for Chinese is probably the lack of significant effect of specificity. Li et al. distinguish between “a specific person or thing that both interlocutors knew” and “one whose referent [bears] general denotation [...] usually represented by *you* or *one* [in English]” (Li et al. 2012: 99), i.e. the subclass “generic” within the factor level non-referential in the present study. They find overt pronoun rates of 47.4% for specific, and 46.3% for non-specific subjects (Li et al. 2012: 104), a result that is in stark contrast to the assessment of generic reference in canonical descriptions of Chinese, and certainly deserves further consideration in future studies (see also section 4.1).

Switch reference is a major factor in Li et al., and a more marginal one in Jia and Bayley. The universal tendency for zero anaphora in contexts of continuous reference is confirmed for all data sets but Indian English, which clearly represents the astonishingly odd exception. On the other hand, a shared insight is the pronoun favouring effect of partial shift, a condition potentially evoking referential ambiguity even more than a full referential switch. The factor level reference

maintenance in the present study conflates overt reference with both identical and shifting form, an approach that is supported by the similar behaviour of both types in Li et al. (2012: 104).

One further structural context needs to be considered: Chinese treats predicative adjectives as verb phrases with overt or omitted subject pronouns (see also section 5.5.4). Including these contexts for the Singlish data, where copula deletion, or alternatively the use of predicative adjectives, is common, would further raise the omission rate for this variety. The sheer numerical contrast between 12% null subjects in Singlish, and roughly 50% in Chinese is thus not as meaningful as it appears at first glance. Both issues prove Torres Cacoullós and Travis' (2014: 22) point that comparing mere omission rates across languages and varieties is not very insightful by itself: first, it needs to be sufficiently transparent how these rates are calculated, afterwards, the conditions of their occurrence need to be clarified in order to identify true parallels in the grammatical systems.

Besides inter-study discrepancies in the categorisation of the data, another issue concerns the kind of data analysed. Acknowledging that the choice of anaphoric form can rely on physical presence of the referent, including telephone conversations, as done by Jia and Bayley (2002), is at least a possible obstacle to comparability with studies analysing face-to-face conversations. Physical presence is one major clue positioning a referent in the category of both familiar and activated entities (see section 2.1.2), and the responsible visual clues are obviously lacking in telephone conversations. The other type of spoken data analysed in both Jia and Bayley (2002) and Li et al. (2012) consists of teacher and student speech, which certainly constitutes a specific speech situation not reliably equivalent to naturally occurring conversation amongst peers. The effect of situational factors and genre on subject realisation is well attested, and also found significant in both Chinese studies. Concerning Jia and Bayley (2002), a further aspect concerning their data deserves mention: they investigate speakers in a Chinese heritage school in the US. This means that the students are actually learners of Chinese rather than native speakers (Jia and Bayley 2002: 106). Regarding the teachers' speech, contact effects from English on NSLs like Spanish are attested, resulting in higher rates of overt subject pronouns in US varieties of Spanish; this is also a possible factor for Jia and Bayley's data, although it is not entirely clear whether the language of the teachers constitutes a true contact variety.

Considering the long list of limitations and caveats provided above, the author feels almost compelled to undertake such a parallel study on Chinese conversational data, but that goes far beyond the resources and scope of the present study. The comparison presented here is thus not conclusive by any means, but rather a first sketch, and decidedly a nudge towards further research.

5.6.4 Conclusion

The comparison of the results achieved by the multivariate analysis of variable subject pronouns reveals common tendencies, but also distinct contrasts among the Asian Englishes, which can be traced back to differences in nativisation status and societal functions of the respective varieties. This confirms earlier observations on the structural consequences of a more advanced nativisation status (e.g. by Schneider 2007), and on the stronger influence of variety type on the morphosyntactic structure of a variety compared to areal patterns (found e.g. by Kortmann and Wolk 2012), even on this very fine-grained level of investigation.

The three Asian varieties differ both in terms of frequency, and in the kinds of null subjects attested: each variety exhibits individual contexts favouring null subjects that the others lack, and that influence the statistical weight of the different factor groups. Several structural predictions from descriptive accounts presented in chapter 4 and comparative surveys like eWAVE (Kortmann and Lunkenheimer 2013) can be confirmed for the data investigated here, e.g. the preference for null subjects in Hong Kong English relative clauses and the non-referential *it* is construction, while the strongly contrasting ratings of dummy subject deletion (pervasive for Indian and Singapore English, attested absence in Hong Kong English) are not as obvious here.

Beyond null subject pronouns, Singapore English especially shows a number of properties associated with topic-prominent languages, such as Chinese-style topics, double NPs, null objects and bare conditionals (see also section 5.5.4). As for null subjects, it is difficult to judge the relevance of these features for claims of a typological shift of the variety without further quantitative evidence for their actual usage, a task that is reserved for future research.

Complementary to the evaluation of substrate effects conducted here, chapter 6 offers the superstrate perspective, providing the comparison with results on ICE-GB from chapter 3.

6 Discussion and Conclusion

6.1 Superstrate comparison: How English are Asian Englishes?

The comparative variationist method proposed by Poplack and Tagliamonte (2001), Tagliamonte (2008), and Meyerhoff (2009) was spelled out in section 4.4.2. It requires the detailed investigation of structural constraints on the linguistic variants in focus. This final chapter concludes the analysis with the comparison of the Asian Englishes with their shared British English superstrate, concerning the frequency and determinants of variable subject pronoun realisation. The analyses in chapters 3 and 5 provide the insights needed to perform this kind of comparison.

6.1.1 Comparative variationist measures: Results

The aim of comparative variationist analysis is to uncover structural parallels and differences in historically related varieties of a language. The Asian varieties discussed here share subject-prominent British English as their superstrate, and topic-prominent languages like Chinese, Hindi, and Malayalam as substrates; their linguistic ecologies are thus similar with regard to the feature under investigation. However, as has been shown in section 5.6, the degree of substrate convergence varies in the different Asian Englishes. The question to be tackled here is whether this divergence from substrate patterns manifests in superstrate convergence, or in innovative patterns of contact languages. The identical set of analytical factors for the superstrate variety, provided by the analysis in chapter 3, facilitates a comparison that is even more comprehensive than the substrate comparison conducted in the previous section, where the depth of comparison is limited by the unavailability of corresponding detailed studies on the substrate languages. Differences between the varieties regarding the variable context of null subjects are discussed in sections 5.1.1 and 5.6.1; the present section addresses the contrastive evaluation of the comparative variationist measures of structural convergence: statistical significance and range of factor groups, i.e. relative factor strength, and the constraint ranking, or hierarchy. Meyerhoff's (2009) taxonomy of transfer is focussed on constraints replicated or calqued from the substrates, and is consequently consulted in section 5.6 for the comparison with the Chinese substrate. The comparative variationist method is based on the same principles and conceptualised more generally; it is thus equally suitable to characterise varying degrees

of superstrate influence in contact varieties. The main interest is the historical development of varieties. The dynamic nature of emerging grammatical systems is incorporated via an evolutionary perspective on dialect formation. According to Tagliamonte,

[t]hese measures [i.e. relative factor strength and constraint hierarchy] enable us to infer whether the data sets under comparison share an underlying grammar, and to what extent. For example, if the constraint ranking of one (or more) factor groups is shared by a set of varieties, we infer that they have inherited it from a common source. If the constraint ranking of factors is parallel, but operates at varying strengths or patterns in different varieties, this can be explained by the *stage of development of the system of grammar* under investigation as represented by each data set. (Tagliamonte 2008: 132–133, emphasis mine)

In addition to these diagnostic measures comparing the effects of single factors, the following comparison evaluates the relative weight of the sub-groups of factors in the different varieties, given the provenance of these different factors (see Table 3.4). Table 6.1 provides the overview of factor group significance, based on the Wald statistics of the respective regression models, and the range of the factor groups, based on the Rbrul centred factor weights in the four varieties investigated. Non-significant factor groups are provided in square brackets, the top three factor groups for each variety are marked in bold.

Table 6.1: Comparison of factor group significance and range

	Factor groups	GB	range	IN	range	HK	range	SG	range
English-specific	Coordination	***	72	***	57	***	55	***	59
	Clause type	*	27	***	54	**	33	***	32
	Position	***	56	—	[19]	—	[18]	***	25
Contact	Person	*	25	***	26	***	32	—	[09]
	Reference	*	20	—	[03]	—	[11]	***	21
Universal	Switch	.	[16]	**	51	***	25	.	[10]
	Persistence	**	35	***	63	***	38	***	23
	Verb type	***	44	***	62	***	36	***	19

The majority of factor groups is significant for all four varieties under investigation. The Asian Englishes show higher statistical significance, and for more factor groups overall compared to British English, i.e. the number of significant constraints on null subjects is higher, but the weight of the individual constraints is lower, especially for Hong Kong and Singapore English. As discussed in section 5.6, Indian and Hong Kong English conform concerning the non-significant factor groups position and specific reference (see also Table 5.47). A more surprising

insight is the parallel structure of British and Singapore English in this regard: both lack significance for the universal factor group switch reference. Person is not significant in Singapore English, and shows only limited influence in British English, especially compared to the strength of the factor group in the Indian and Hong Kong varieties. Judging from factor group significance alone, Indian English and Hong Kong English are clearly similar, and contrast notably with the two L1 varieties. Following Tagliamonte’s argument, this provides evidence for similar stages of development in the grammatical systems of the two L2 varieties, an interpretation that is in line with the comparable status of English in the two speech communities as a tool for communication, rather than taking the form of a highly indigenised marker of local identity.

Regarding the range of the factor groups, while most factors are statistically significant in Singapore English according to their Wald statistics (see Table 5.35), throughout the different factor groups the range is consistently lower than for the other varieties. The lower range indicates lower contrasts among the different factor levels, and hints towards the lesser explanatory value of the regression model for this variety, which is also evident from the diagnostics of model significance and validity discussed in section 5.5.2. This shows the importance of incorporating the factor range into the evaluation of factor effects beyond the mere values of significance. At first sight, the model seems to work very well in explaining Singapore English null subjects, but even the most influential factors in Singapore English have ranges comparable to, or even lower than, the least influential factors in the other varieties.

A comparison of the constraint rankings provides further evidence for similarities or differences in the underlying grammatical systems of varieties (Tagliamonte 2006: 245). Additionally, the relative position of English-specific, contact, and universal factors within these rankings can point towards the origins of structural constraints, and the relevance of these different sources and mechanisms of grammatical evolution for the individual grammatical systems. Table 6.2 presents the constraint rankings provided by the Rbrul output for logistic regression of all four varieties order of decreasing strength. The different factor subgroups are marked as **bold** = English-specific, *italics* = contact, underlined = universal, non-significant factor groups are given in square brackets.

Table 6.2: Comparison of Rbrul constraint hierarchy

GB	IN	HK	SG
Coordination	<u>Verb type</u>	Coordination	Coordination
<u>Verb type</u>	Persistence	<u>Verb type</u>	<u>Verb type</u>

Table 6.2: (continued)

GB	IN	HK	SG
Position	Coordination	<i>Person</i>	<i>Specific reference</i>
<i>Person</i>	Clause	<u>Switch reference</u>	Position
<i>Specific reference</i>	<u>Switch reference</u>	<u>Persistence</u>	Clause
<u>Persistence</u>	<i>Person</i>	Clause	<u>Persistence</u>
Clause	[Position]	[Position]	<u>[Switch reference]</u>
<u>[Switch reference]</u>	<u>[Specific reference]</u>	<u>[Specific reference]</u>	<u>[Person]</u>

A comparison of the rankings provides evidence for the universal effect of verb type on subject omission: the factor group is ranked high, i.e. second or first, in all varieties. The shared constraint within this factor group is the blocking effect of primary auxiliaries on subject omission, and the favouring tendency of the least specific class of lexical verbs, whereas psychological verbs tend to have lower omission rates. There is no consensus on the behaviour of modal auxiliaries: for Hong Kong and Indian English, they show similarly low omission rates to primary auxiliaries, in British and especially Singapore English they permit higher rates of null subjects (see also Figure 6.1 below). Whether this is due to the influence of individual lexemes (see Figure 3.7 and Figure 5.7), or can be accounted for by the higher syntactic complexity of verb phrases containing modal verbs remains a question for future research (see e.g. Wagner 2016 on the effects of verb phrase complexity in Newfoundland English).

A wide-reaching effect is also observed, if to a slightly lesser degree, for the English-specific factor group coordination, which is the top-ranking constraint in British, Hong Kong and Singapore English, but only ranked third most influential in Indian English. As postulated by Torres Cacoullos and Travis (2015), and addressed in section 3.6, the effect of coreferential coordination on subject omission is possibly not as characteristic for English as usually claimed in the literature. Instead, it provides a context that favours pronoun omission from an information processing perspective based on universal cognitive processes rather than language specific patterns.

The opposite is shown for the supposedly universal factor switch reference, which is not statistically significant in British and Singapore English, and plays a comparatively limited role in the L2 varieties. While referential continuity is integrated into the factor coreferential coordination by design, beyond this rather specific context continuous reference alone is not sufficient to grant predictability of the omitted form and significantly favour zero anaphora. This confirms findings by Wagner (2012), who tackles persistence effects by combining preceding token, turn boundary and switch reference, but finds that immediately

preceding zero tokens alone are by far the most influential predictor for further subject omission. Immediately preceding null subjects are also influential in the present study, especially for the L2 varieties Indian and Hong Kong English. It is shown for all varieties that this effect is independent from continuous reference (see Figure 3.13, Figure 5.13 and Figure 5.18). Clusters of null subjects are thus not necessarily related to topic chains indicating coherent reference across stretches of discourse as they are typical for the topic-prominent substrate languages, but rather reflect a more universal tendency for mechanical priming also observed by Torres Cacoullós and Travis (2014).

The behaviour of the other factor groups seems more haphazard across varieties. The factor group person shows no clear tendency, and especially the presumed contact factor specific reference behaves counterintuitively: while slight statistical significance for specific reference is attested for British English, where no such effect was expected, the complete lack of effect in Indian and Hong Kong English, and the limitations of the factor beyond the calqued *got*-existential construction in Singapore English is surprising. Even more than anaphoric pronouns with specific reference, non-referential pronouns seem to be bound to specific collocates, such as *seems*, *means*, and Hong Kong English *it is*. The complete elimination of non-referential null subjects is also attested in the development of South African Indian English basilect, mesolect and acrolect in Mesthrie's comparison of null subject rates (see section 4.2), and mirrors the progress of overt pronouns in L1 acquisition.

Structural conflict sites are a crucial element in the comparative variationist analysis (as addressed by e.g. Poplack and Meechan 1998, Poplack and Tagliamonte 2001, Torres Cacoullós and Travis 2015). For the investigation of contact effects on null subjects, Torres Cacoullós and Travis (2015) propose the restriction to initial position as a conflict site between English and the canonical NSL Spanish. There is no evidence for this position constraint from the descriptions and analyses of topic-prominent languages (see section 4.1). The same conflict site can thus be assumed for contact between English and the Asian languages discussed here. Indeed, position is clearly less significant for Indian and Hong Kong English. British English shows an especially marked rejection of subject omission in second position that is not found in the contact varieties. The weakening of the position constraint in the Asian Englishes can be interpreted as a clear contact effect on a superstrate-specific structural constraint.

A further possible conflict site is the factor group clause type. The almost categorical restriction of English null subjects to declarative main clauses is not valid to the same degree for the Asian varieties; especially Hong Kong English and Singapore English show a marked deviation in variety-specific contexts. Obviously, in English questions and subordinate clauses subject pronouns are

also less likely to be in initial position. The possible relation of these two factor groups thus calls for further research.

6.1.2 Comparison: First person pronouns

Different grammatical persons are widely assumed to show divergent patterns of variation for subject realisation. Analogous to section 3.6.2, this assumption is tested for the Asian varieties, although the lack of comparable substrate studies restricts the possibilities of contrastive analysis. The results provided by the analysis of ICE-GB (see section 3.5) serve as comparative foil instead. Due to the limited token numbers of first person contexts, contrasts described here cannot be stated conclusively, but are rather intended as a presentation of first tendencies, awaiting further research.

As a further obstacle to comparability, the behaviour of the factor group person is notably different in the three varieties (see also Figure 5.26). Person is highly significant for Indian and Hong Kong, but not statistically significant for Singapore English, which shows little variation between persons concerning subject omission rates in most contexts. In fact, first person is the most common context for null subjects in Singapore English by a small margin, while it is least likely to be omitted in Indian English, and even more restricted in frequency in the Hong Kong data. Again, Singapore English in this respect is more similar to the superstrate than to either Asian variety.

The clause type question and the verb type modal are not attested in Indian and Hong Kong English for first person contexts. Singapore English does exhibit first person zero in these environments, although they are few and far between. Table 6.3 shows a contrastive overview of the statistical significance of the relevant factor groups in the full vs the first person data sets of the Asian varieties, with values from ICE-GB as reference.

Table 6.3: Comparison of factor group significance for first person contexts vs full data set

Factor groups	GB 1st	GB all	IN 1st	IN all	HK 1st	HK all	SG 1st	SG all
Coordination	***	***	***	***	***	***	***	***
Clause type	.	*	—	***	.	**	—	***
Position	—	***	—	—	—	—	**	***
Switch	—	.	**	***	***	***	—	.
Persistence	***	**	***	***	***	***	***	***
Verb type	**	***	**	***	**	***	**	**

Less factor groups are significant for the reduced data sets, which is at least partly due to the lower token numbers in the different contexts. Compared to the full data set, the clearest loss of statistical significance is exhibited for the factor group clause type, which is limited to the binary distinction main vs subordinate clause in ICE-GB, ICE-IN, and ICE-HK. While first person null subjects in subordinate clauses are exceedingly rare, the factor group is not statistically significant in any of the statistical models for first person contexts.

Apart from the more restricted variable context for first person null subjects addressed above, the other factor groups behave almost the same as in the full data sets: switch reference is not significant for British and Singapore English, position is not significant for Indian and Hong Kong English, and additionally loses significance in British English. Both the subgroups of significant factors and their ranking remain largely equal across varieties compared to the full data set. One notable contrast is the larger role persistence apparently plays for first person contexts than for the full set. This is possibly explained by the more diverse functions of second and third person pronouns. First person pronouns invariably include speaker reference; this meaning even persists to a degree in generic uses of the first person plural, referring to “(us) people in general”. This semantic component is less distinct in second person generic pronouns, and completely absent from third person, both with specific and non-specific reference. Referential coherence is thus more pronounced and unambiguous in clusters of first person pronouns, overt and zero.

Verb type is possibly the most interesting factor group for first person contexts, considering the reported tendency for overt first person pronouns with psychological verbs. Again, it is Singapore and British English that show parallel effects, with the likelihood for null subjects decreasing in first person contexts for psychological verbs (Figure 6.2), compared to the full set including all persons (Figure 6.1). Indeed, psychological verbs have a lower likelihood for subject omission in first person contexts compared to the full set.

First person pronouns behave very differently across the three Asian Englishes, which makes a conclusive evaluation of this context difficult. Still, the broader variable context for Singapore English first person pronouns compared to the other varieties adds further evidence for the shift from minor to (comparatively) major use pattern of null subjects in this variety. In conclusion, the analysis of subsets of subject pronouns possibly leads to misled generalisations, especially visible in the different envelopes of variation addressed in section 3.1 for British English, and in section 5.1 for the Asian Englishes, and the diverging omission rates between persons in the Asian varieties.

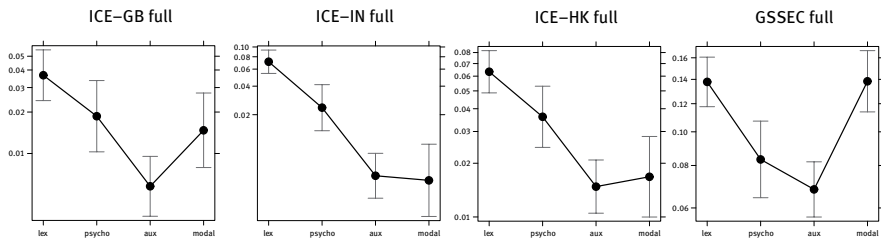


Figure 6.1: Comparison effects verb type full data set

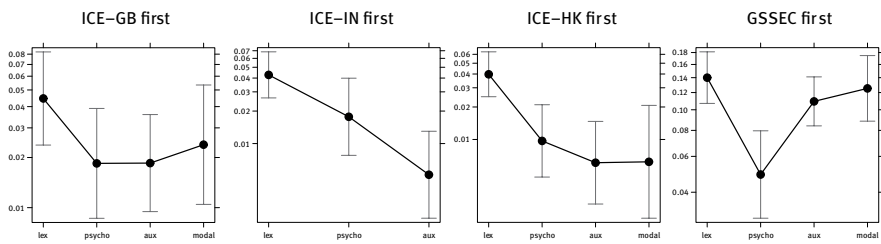


Figure 6.2: Comparison effects verb type first person

6.1.3 Summary

Besides factor group significance (Table 6.1) and constraint hierarchy (Table 6.2), a further criterion to measure the similarity of grammatical systems is the order of factor levels within factor groups, indicating the relative favouring or blocking effect of the respective structural context on the response variant. The order of factor levels is based on decreasing factor weights measured on the probability scale, with values >0.5 favouring omission, values <0.5 favouring overt pronouns, and values close to 0.5 indicating a lack of effect for the respective factor level. This order is represented in the tables for individual factor groups in sections 3.4.3, 5.3.2, 5.4.2 and 5.5.2. An alternative way of representing the order of factors within factor groups is the graphic representation of the direction of effects in the effects plots (see also Figure 3.15, Figure 5.16, Figure 5.21 and Figure 5.24). In contrast to the figures presented above, the plots shown in Figure 6.3 are scaled to adjust for the different range of the factor groups in the individual varieties and allow for direct comparison.

The synoptic graph shows that across all categories, Singapore English has a higher probability for null pronouns, visible from the consistently higher level

of the blue line. The relative weakness of the predictors for Singapore English null subjects is also visible from the lack of strong amplitudes for most categories, which reflects the low range of the factor groups for this variety. Differences between factor levels are stronger in the other three varieties, but especially British English and Indian English. The least significant factor group across varieties is obviously specific reference, where all four lines are almost horizontal. The direction of effects is the same, i.e. the lines run largely parallel for the categories coordination, switch and persistence, while the differences addressed above for the categories clause, position, person and verb type are nicely illustrated by the intersecting lines.

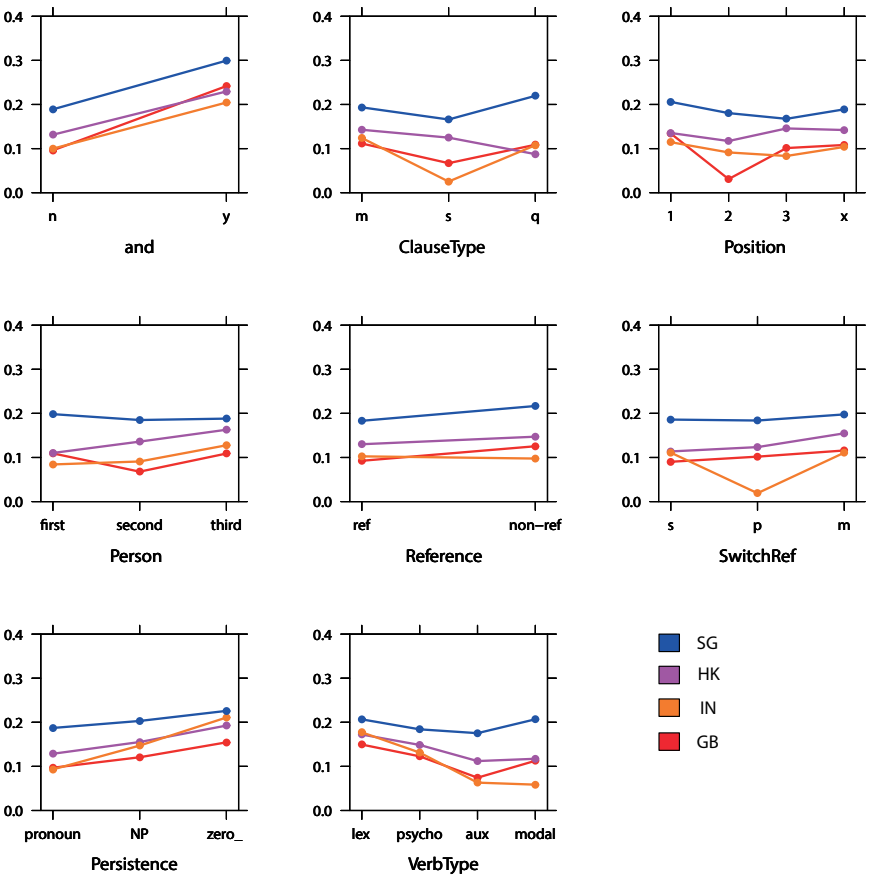


Figure 6.3: Comparison predicted probabilities

6.2 Variable rules, variationist typology and language contact

The results of the present study are relevant for two different aspects of further linguistic investigation. The canonical syntactic configuration of English requires invariant overt subject pronouns, but this categorical claim is qualified by the results presented here. What are theoretical and methodological implications of the regular occurrence of null subject pronouns in a prototypical non-NSL, as described in chapter 3? On the other hand, contact between languages at opposite ends of an assumed “null subject scale” leaves measurable quantitative traces in the grammatical system of the respective contact varieties, which makes this phenomenon a suitable candidate for a comparative multivariate analysis within the variationist paradigm. As a case study of contact-induced grammatical change, the investigation of null subjects in Asian Englishes conducted in chapter 5 tackles the question of how the variationist method can contribute to the study of contact varieties of English by supplementing theories of language contact with detailed empirical evidence. These issues are discussed in turn in 6.2.1 and 6.2.2, providing an evaluation of the results of the present study and sketching possible future research avenues.

6.2.1 Null subjects in non-NSL: Theoretical and methodological implications

As the first study of this scope, the present investigation is a contribution towards a better understanding of null subjects in a non-NSL like English, an issue that has only very recently received considerable attention from a variationist perspective. The generative position on null subjects encountered in non-NSLs is that they are either performance errors, or of a completely different type than canonical null subjects. Both non-canonical NSLs and non-NSLs have attracted fresh linguistic perspectives in recent years, calling into question the original binary conception of the null subject parameter (see sections 2.1, 4.1 and 4.2). The incipient variationist analysis of null subjects in traditional L1 varieties of English, especially in various studies by Rena Torres Cacoullós, Catherine Travis, and Susanne Wagner, depends on the assumption that even in presumed non-NSLs, null subjects do make a regular, if rare, appearance, and are systematically conditioned by factors similar to those established for canonical NSLs (see section 2.3). Based on the diagnostics of model validity for the logistic regression model, variable subject realisation in British English can be well explained by multivariate analysis (see section 3.4). This indication for the systematic nature of subject omission in British English contributes additional evidence for the gradual nature of supposedly discrete typological parameters.

Comparability between studies is indispensable for scientific progress, and crucially depends on transparent scientific practice. The variationist method adopted here relies on descriptive approaches to grammatical systems and painstaking scrutiny of the data, complemented by the researcher's intuition on the differentiation between meaningful and more peripheral differences. Adopting a comparative perspective adds further challenges to the crucial analytical stages of data categorisation and coding. The variable context in the present study is circumscribed rather generously in order to account for substrate structures in the contact varieties that deviate from the Standard English system, and act as a baseline for cross-variety comparison (see section 3.1). While this strategy is necessary for the comparative analysis, it naturally falls short in producing the best possible explanatory model for each individual variety (related problems of fine-grained cross-linguistic comparison are addressed in Kortmann 2014).

Different designs of the variable context in analyses discussed at various points in the present study lead to differences in subject omission rates that paint a considerably skewed picture, be it through the inclusion of imperatives and predicative adjectives in the Chinese studies (see also section 4.1), or the artificial data reduction conducted for most studies on English null subjects (see also section 2.3). The consequences of this procedure for the following statistical analyses are unclear, especially considering Wagner's (2012: 131–133) discussion of the behaviour of GoldVarb, the statistical tool of choice for most preceding studies, with low-frequency linguistic contexts. Furthermore, taking the role of individuals in language variation seriously, which is increasingly implemented by the integration of speaker as a random variable in mixed regression models, entails the exclusion of subject tokens by invariant speakers from the analysis, an aspect that is not addressed in Torres Cacoullós and Travis' (2014) study, and apparently not relevant for Wagner's (2012, 2016) data. Judging from the results of the present study, another aspect worth further attention is the role of individual verb tokens or collexemes in subject omission. While verb tokens as random effects are becoming an established part of variationist analyses of morpho-syntactic variables, the low frequency of null subjects in English precludes the straightforward adoption of this approach. Moreover, the widespread practice of artificially reducing overt pronominal subject tokens completely obliterates the possibility to make quantitative statements on the influence of collocational patterns in earlier studies. How to achieve the proper balance between necessary economy and sufficient detail in the quantitative investigation of low-frequency phenomena is thus a methodological issue yet to be resolved.

The present study has been successful in showing that English null subjects are subjected to variation governed by variable rules, which exhibit intra-linguistic variability between the varieties analysed here. In the spirit of Torres Cacoullós

and Travis' (forthc.) *variationist typology*, the cross-linguistic significance of these findings is discussed in the following. The strong effect of coreferential coordination for all four varieties studied here strengthens Torres Cacoullos and Travis' (2015) doubts about the English-specific status of this context. It is intuitively appealing to assume a more general validity of this constraint on a cognitive basis – coordination, especially when overtly marked by a conjunction, links two (or more) predicates of the same subject in a contiguous form. It commonly implies situational continuity, making the referent of the zero pronoun highly predictable. However, when considering the effects of coreferential coordination, it is crucial to make the circumscription of this context transparent: is it defined in broader conceptual terms, including e.g. lists and asyndetic coordination, or confined to those cases marked more explicitly by conjunctions like *or* and *but*, or in the most narrow sense employed by Torres Cacoullos and Travis (2014) and the present study, restricted to the schematic [PRONOUN_i VERB *and* Ø_i VERB]?

While coordination is thus a possible candidate for promotion from language specific to universal constraint, the opposite tendency is observed for the assumed universal factor switch reference. Principles of information processing are commonly evoked as an explanation for this factor. However, the effect of null anaphora on processing complexity is assessed differently. Lezama and Almor (2011) show that non-emphatic overt pronouns and repeated names, i.e. redundant anaphoric elements, cause a processing delay for the canonical NSL Spanish, but the lack of this effect for Chinese calls into question the universal nature of this “repeated name penalty” and is more in line with C.-T. Huang's (1984) evaluation of null anaphors as additional interpretational load on the hearer. Like previous studies on English, the present one has failed to establish statistically significant effects for the factor group switch reference outside coordination. While a higher likelihood of null anaphora in contexts of continuous reference is an immediately plausible prediction, the validity of this claim beyond canonical NSLs is not proven beyond reasonable doubt. These inflectional languages, however, profit from the related factor “discourse connectedness”, manifested in continuous morphological marking. Persistence, on the other hand, is not as established as switch reference in the study of subject pronoun omission, but seems a likely candidate for a universal structural constraint based on general psycholinguistic principles such as priming, and thus deserves further cross-linguistic investigation.

Processing load is also evoked in Wagner's (2012, 2016) discussion of VP complexity. This factor represents an attempt to operationalise and explain more anecdotal observations on the higher frequency of subject omission on past, and especially negated verb forms. In the present study, the better known and obviously related factor verb type takes precedence, together with the evalua-

tion of collocational patterns; however, the different status of learner features in the Asian varieties, depending on their differences in nativisation, makes VP complexity a promising factor for future studies on subject omission in contact varieties. Considering their cognitive basis, especially the factor groups associated with processing load call for experimental approaches complementing the analysis of corpus data, e.g. measuring comprehension rates of different structures (Tao and Healy 2005), but also the targeted evocation of production data for different linguistic contexts (see e.g. Chen and Pan 2009, Li 2014, Xiao 2002).

Intonation is a factor neglected by the present study. Its role in marking given and new information in spoken language is crucial, most notably here in the distinction between stressed (strong) and unstressed (weak) pronouns as a further dimension of accessibility marking according to Givón's topicality hierarchy (Givón 1983; see also section 2.1.2; for a more detailed assessment of stressed and unstressed subject pronouns in English, see Travis and Torres Cacoullos 2014). Moreover, differences in intonational patterns form a major dividing line between speakers of different ethnic backgrounds in Singlish (Lim 2001), their investigation thus possibly leads to different groupings within and among the data sets analysed here. Complementary investigation of more general contrasts in information management of different language types and varieties, incorporating lexical, syntactic and phonological aspects, are a vivid field of study (see e.g. Calhoun 2010, Heusinger 1999), and promise valuable insight for the study of contact varieties of English as well, especially those with tonal substrate languages.

A central aspect of variationist investigation is the role of stylistic variation. As for many sociolinguistic variables, text type is crucial for the realisation of subject pronouns, which makes a comparison with results based on different types of data difficult. The present study is focussed on direct conversation, where the physical presence of a referent is one major clue for its accessibility (section 2.1.2); the absence of shared physical context in telephone conversations hinders comparability with the results of Jia and Bayley (2002) for Mandarin Chinese. On the other hand, teacher-student classroom interaction also constitutes a very specific communicative situation. This is confirmed by both Jia and Bayley (2002) and Li et al. (2012), who find widely varying deletion rates for different persons in the different text types. These differences are presumably based on the predictability of certain forms in the *specific* situational context. Second person singular is unexpected in the classroom setting, where the teacher is more likely to address the group as a whole and second person singular thus exhibits a higher rate of overt pronouns, whereas second person plural reference is unexpected in telephone conversation, and consequently is more likely to be expressed in this text type. Even *within* specific contexts, studies on Singlish in classroom contexts

show that there is linguistic variation on almost microscopic levels of situational- or topic-shift within a session, of both teachers and students, according to task and content (see also section 5.5.4).

Narratives constitute another specific text type, relevant here for their role in the study of information structure (e.g. the Pear Stories). Especially the relatively free conversations in the GSSEC feature several narrative stretches within the conversational setting. As shown by Travis and Lindstrom (2016), in the English Pear Stories, null subjects are significantly more frequent than in Standard American English conversation (as represented by the SBCSAE). Especially the rate of null subjects found for third person contexts is higher in narratives, which typically show a high degree of topic- and character-continuity. Moreover, TAM marking of the verb, a significant factor in studies on Spanish null subjects, possibly manifests differently in linear narratives exhibiting more contiguous temporal sequentiality than conversations. Judging from the results on Chinese classroom speech, and English narratives, person is one structural constraint decisively affected by the text type analysed. Studies restricting their investigation to one person thus need to proceed carefully when drawing more general conclusions on the distribution of null subjects.

Through their anaphoric function, the usage of referential pronouns is necessarily bound to the content as well as the context of speech. Given the availability of additional genres in the ICE corpora, a logical next step is the inclusion of further sub-corpora of the respective varieties in order to test Travis' (2007) and Travis and Lindstrom's (2016) hypothesis on genre variation as a purely quantitative phenomenon that leaves underlying structural constraints untouched. This is possibly not true to the same degree in World Englishes, where issues of norm orientation and codification might lead to rather different outcomes between different types of speech.

Further aspects of language contact and linguistic nativisation for subject pronoun expression are discussed in the following section.

6.2.2 Language contact and linguistic nativisation

The stylistic dimension and the indexical use of non-standard features is of central relevance for the study of World Englishes. Higher amounts of idiosyncratic behaviour of individual speakers are observed for the L1 varieties British and Singapore English, compared to the more homogeneous behaviour of the L2 varieties Indian and Hong Kong English. Given the function of English in India as a non-native link language, and the primary use in ICE-HK as a means of communication with non-Hong Kong speakers, the degree of informality encountered in

the conversations in ICE-GB and GSSEC is probably not even available for the L2 varieties, an assessment shared by Percillier (2016: 192). While null subjects are considered an informal, colloquial feature in Standard English, their status in the Asian varieties is not as clear, not least due to their association with substrate origins. On the other hand, Schröter (2010) shows that variation between formal and informal speech in Singapore English is marginal for subject omission; the feature is apparently below conscious usage, and not employed as a sociolinguistic marker the way e.g. BE deletion is. The comparative analysis presented here provides the first steps in the systematic description and structural circumscription of variable subject pronouns in Asian varieties of English. Its results constitute a good basis for the design of experimental tasks further exploring different structural constraints, such as acceptability ratings, comprehension tests, or syntactic pattern completion, that specifically target structural conflict sites of language contact, and possibly include speakers of different demographic backgrounds (see e.g. Jin 1994, Spring and Horie 2013, Yuan 1997).

Regarding the frequency of subject pronoun omission, the different degree of cultural indigenisation in the three Asian Englishes investigated here is mirrored by the amount of linguistic nativisation. This provides further evidence for the precedence of variety type over the areal effect of shared substrates on the morphosyntax of contact varieties. In principle, the three Asian varieties should exhibit similar feature pools inherited from their topic-prominent substrates, from which they draw structural features. It is thus quite likely that pronoun drop is actuated in all three Asian varieties studied here, but propagated only in Singapore English (see e.g. Lange 2012: 236). What determines the fate of a given morphosyntactic feature as either an ephemeral interlanguage phenomenon, or as a lasting characteristic of an indigenised variety is rather impossible to predict; however, favouring factors both from linguistic and extralinguistic perspectives can be identified post hoc. These are discussed in the following.

According to Schneider, the most fertile ground for structural change in contact varieties of English lies at the lexico-grammar interface (Schneider 2007: 46). This is also addressed in Mukherjee and Gries' (2009) discussion of collostructional nativisation. Apparently, this process does not only apply to complementation patterns and transitivity, but also to specific lexical contexts for the contact feature subject omission in the Asian Englishes. Verb type is identified as highly influential for all varieties in this study. The influence of this factor manifests in two ways: verb types in terms of semantic and syntactic features, and the influence of individual collocates in the different varieties. Torres Cacoullos and Travis (2014) assume the cross-linguistic presence of lexically specific constructions favouring subject omission, but expect them to manifest with different, language specific collocates. The results of the present study corroborate this

assumption: each variety shows idiosyncratic collocates of null subjects, such as British English *sounds* and *see*, Indian English *means* and *asked*, Hong Kong English *is*, and Singapore English *depends* and *can/cannot*. In cases like the Singlish discourse particle *know*, Wee (2003a) claims that its emergence is facilitated by the status of Singlish as NSL. Reversely, the presence of idioms like *cannot* provides a nurturing context for pronoun omission, and a possible starting ground for further context expansion.

One stronghold of null subjects in Singapore English, expletive subjects specifically, is the existential *got*-construction. This structure is classified as grammatical replication by Heine and Kuteva (2005). Its propagation is supported by the relatively strong attraction of *get* and *got* (with possessive and modal meaning) to null subjects in British English (see Table 3.1). Moreover, existential *got* follows a cross-linguistically observed path of grammaticalisation; consequently, its use is also attested Malaysian English (Percillier 2016: 93). This convergence of cross-linguistic developments with suitable lexical exponence in the lexifier is the backbone of Bao's (2010) lexifier filter model, which tries to account for the varying results of this selection process post-hoc, and calls for quantitative investigation like the analysis provided here for more nuanced insights into the consequences of this filtering mechanism.

Chapter 5 also shows the limits of comparative variationist analysis. While Hong Kong English and especially Indian English can usefully be measured against the same analytical categories as their superstrate, the predictive value of the explanatory structural factors is rather limited for Singapore English. In Singapore English, the topic-prominent Sinitic languages and related substrates of the same language type, like the contact languages Baba Malay and Bazaar Malay, have led to structural changes so extensive that their effect is evaluated as a typological shift from subject prominence to topic prominence (e.g. by Ansaldo 2004, 2009, Bao and Lye 2005), going as far as Ansaldo's claim that Singlish "is more an Asian variety with English influences than a variety of English" (Ansaldo 2009: 145). It is remarkable that this shift is so visible in the informal educated variety of Singapore English, which is the basis for the present investigation, whereas so far this fundamental transformation has mainly been claimed for the basilectal variety Singlish. On the one hand, as section 5.5.4 shows, this influence stretches beyond the rather rigid analytical categories necessary for systematic multivariate analysis. On the other hand, subject deletion rates in the educated variety are not as remote from Chinese as it seems at first glance: the approximately 50% null subjects reported for Mandarin Chinese in detailed quantitative investigation by Jia and Bayley (2002) and Li et al. (2012) include imperatives and predicative adjectives, manifesting as BE deletion in Singapore English. In a previous study on Singapore English, Schröter (2010) takes contexts of null subjects

without overt BE into account as well. This wider envelope of variation almost doubles the omission rate in Singapore English conversations to more than 20% of pronominal subject tokens, compared to 12% in the present study. A further indicator for topic prominence is the amount of null expletives, a feature usually associated with lower degrees of linguistic proficiency in earlier stages of language learning (see e.g. Phinney 1987). In generative approaches, overt expletive pronouns in the target language are deemed a crucial hint for parameter setting in language acquisition; their presence unambiguously signals the setting [- pro-drop] and triggers its entailing structural consequences. Referential pronouns, on the other hand are variable in their realisation in NSLs, and thus not as compatible with the conceptualisation of a binary parameter set by cues in the input for both L1 and L2 learners. However, the role of null expletives in Singapore English is qualified by their strong attraction to a lexically specific context, the existential *got*-construction. Obviously collocational patterns and preferences provide niches for substrate features, but whether these are the last resort, or are paving the way for future expansion of a substrate feature, is not deducible from synchronic investigation, but requires diachronic data.

The contrasts between the Asian varieties analysed here align with their position in different phases of Schneider's evolutionary cycle of postcolonial Englishes. This opens two further avenues of research: on the one hand, the comparison with speakers of Malaysian English promises more insight on the role of different substrates in Singapore English. While Malaysian English is closely related historically to Singapore English, their developments diverge after Singapore's independence, leaving them at different stages of nativisation (Schneider 2007: 144–160). On the other hand, a comparison of the results achieved here either with historical data of Singapore English, as represented e.g. in oral history interviews by the National Archives of Singapore, or with current acquisition data of L1 speakers of Singapore English, as suggested by Buschfeld (2015, 2018), could enlighten the evolution of variable subject realisation along the process of nativisation in the variety most progressed on this path.

Whether English is used as L1 or L2 in a speech community crucially depends on language-external factors, and its status consequently affects the linguistic system, especially regarding the propagation of innovative structural features in World Englishes. An example for a pattern that should be attested in two of the feature pools, but only is propagated in one of the varieties, is the *X or not* question, which frequently omits subject pronouns. It stems from a variety of Chinese (Teochew) present in both Singapore and Hong Kong, but is only attested in the Singaporean variety of English. The absence of *X or not* questions in Hong Kong English can be explained by the local lack of prestige of the Chinese variety predominantly employing this structure. The sociolinguistic situation is different in

Singapore, and consequently this specific substrate structure is more acceptable in the respective variety. Such explanations are not as evident for null relative pronouns or the existential *got*-construction; for these structures, no differences between varieties of Chinese are attested.

Matras identifies two crucial factors for the propagation of structural innovations in bi- or multilingual communities, namely the “directionality of bilingualism, and the extent of control and pressure that is exerted on speakers to conform to more established speech norms” (Matras 2009: 312).

Although in Singapore there is strong institutionalised pressure towards Standard English, e.g. in the form of the *Speak Good English Movement* (goodenglish.org.sg), there are several aspects that hinder the eradication of features like subject pronoun omission from informal educated Singapore English:

1. The central role of Singlish in Singaporean identity formation gives way to countermovements and linguistic liberation attempts, e.g. the *Coxford Singlish Dictionary* (www.talkingcock.com), the *Speak Good Singlish Movement*, etc., which possibly override prescriptive efforts (see e.g. Ho and Alsagoff 1998, Ooi 2001, Tan 2017, Wee 2014, 2018).
2. Null subjects are not a sociolinguistic marker; there is no evidence of a conscious targeting of this structure from the material presented on the SGEM website, as is the case e.g. for prepositions and morphological marking.
3. Several constructions favouring null subjects in Singlish either fit an existing paradigm, such as the discourse marker *know*, or have an iconic function related to local culture, such as *can/cannot* (Wee 2003, Wong 2015).

This tension between “control and pressure” (Matras 2009) and the urge of speech communities towards linguistic emancipation (Schneider 2007) is characteristic for the evolution of World Englishes and makes Singapore English such a fascinating field of sociolinguistic research.

6.3 Conclusion

The present investigation constitutes a case study of the possibilities of comparative variationist research on World Englishes, using the classic syntactic parameter of variable subject pronoun realisation as a measure of substrate convergence in three contact varieties of English in Asia.

In an ideal world, comparative analysis operates on the basis of investigating identical structural constraints, within consistently defined variable contexts for all varieties or languages under scrutiny. Given different research foci, this is all but impossible without a concerted effort. The difficulty lies in finding the

balance between generalisability of results, as is desirable in comparative, possibly even typological enterprises, and abstracted categories that still appropriately describe the linguistic realities encountered in the respective data sets (see Kortmann 2014). In its broad inclusion of linguistic contexts, the analysis of British English in chapter 3 is intended not only as a comparative baseline for the present study, but will hopefully also be useful as a yardstick of null subjects in spoken Standard English for future research. The surprising surge of empirical research on null subjects in non-NSLs like English in recent years proves the demand for data-based analyses of this low-frequency syntactic phenomenon; the relevance of this approach lies in the systematisation of structural descriptions as a test of anecdotal or impressionistic observations.

Torres Cacoullos and Travis (2015) argue that the analysis of intra-linguistic variability helps to assess “inter-linguistic (dis)similarity”, i.e. the concrete structural presentation and conditions of variables like subject pronouns in different varieties of a language can provide insight into the cross-linguistic validity of structural constraints on subject expression and thus identify truly universal tendencies. I argue that the reverse is also true, and contributes a crucial aspect to the investigation of World Englishes: integrating the typological perspective in the study of language-internal or micro-variation can help to assess the source and status of a feature in a given (contact) variety (see also Brunner 2017, Siemund 2016). This approach is especially fruitful when contact effects between languages of different typological configuration are assessed. By focusing on contact between non-NSL English and topic-prominent, or radical NSLs, the present study complements Torres Cacoullos and Travis’ focus on contact between English and the canonical NSL Spanish.

The preceding analyses in chapter 5 and section 6.1 show that contact with typologically different languages leaves measurable traces in the grammar of the Asian Englishes that can insightfully be analysed by multivariate analysis. Substrate influence in the educated variants of the respective varieties goes beyond the superficial raising of subject omission rates and affects the variable rules determining omission, indicating changes in the underlying grammatical system. Juxtaposing the constraints found for the Asian varieties with those found in their substrates clarifies the contrasts both between substrate and contact varieties, and between the different contact varieties.

To conclude, this study has shown both promise and pitfalls of comparative quantitative investigation of contact varieties. Systematic divergence in the grammar of the Asian Englishes can be identified in quantitative terms, and degrees of contact can be measured using comparative variationist diagnostics to evaluate the relative degrees of language specific and universal tendencies. In the present case, the existence of specific constructions associated with null

subjects shows how lexically specific constructions provide a nurturing habitat for contact features, but whether null subjects are a stable feature of Asian Englishes, are on the retreat, or even on the rise, remains to be seen in future studies.

References

- Alsagoff, Lubna. 2010. English in Singapore: Culture, capital and identity in linguistic variation. *World Englishes* 29(3). 336–348.
- Alsagoff, Lubna & Ho Chee Lick. 1998. The relative clause in colloquial Singapore English. *World Englishes* 17(2). 127–138.
- Anjaria, Ulka (ed.). 2015. *A history of the Indian novel in English*. New York: Cambridge University Press.
- Ansaldo, Umberto. 2004. The evolution of Singapore English: Finding the matrix. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 127–150. Amsterdam: Benjamins.
- Ansaldo, Umberto. 2009. The Asian typology of English: Theoretical and methodological considerations. *English World-Wide* 30(2). 133–148.
- Ansaldo, Umberto. 2010. Contact and Asian varieties of English. In Raymond Hickey (ed.), *The handbook of language contact*, 498–517. Malden, MA: Wiley-Blackwell.
- Anthony, Laurence. 2014. AntConc (Version 3.4.3w). <http://www.laurenceanthony.net/software/antconc/>, accessed on 2019-05-06.
- Ariel, Mira. 1990. *Accessing noun-phrase antecedents*. London & New York: Routledge.
- Ariel, Mira. 1994. Interpreting anaphoric expressions: A cognitive versus a pragmatic approach. *Journal of Linguistics* 30(1). 3–42.
- Aye, Khin Khin. 2006. *Bazaar Malay: History, grammar and contact*. Singapore: National University of Singapore Ph.D. thesis.
- Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge [etc.]: Cambridge University Press.
- Bailey, Laura. 2011. Null subjects in Northeast English. *Newcastle Working Papers in Linguistics* 17. 23–45.
- Balasubramanian, Chandrika. 2009. *Register variation in Indian English*. Amsterdam & Philadelphia: John Benjamins.
- Bao Zhiming. 2005. The aspectual system of Singapore English and the systemic substratist explanation. *Journal of Linguistics* 41(2). 237–267.
- Bao, Zhiming. 2009. One in Singapore English. *Studies in Language* 33(2). 338–365.
- Bao, Zhiming. 2010. A usage-based approach to substratum transfer: The case of four unproductive features in Singapore English. *Language* 86(4). 792–820.
- Bao, Zhiming. 2012. Substratum transfer targets grammatical system. *Journal of Linguistics* 48(2). 479–482.
- Bao, Zhiming. 2015. *The making of vernacular Singapore English: System, transfer and filter*. Cambridge & New York: Cambridge University Press.
- Bao, Zhiming & Khin Khin Aye. 2012. Bazaar Malay topics. In Umberto Ansaldo (ed.), *Pidgins and Creoles in Asia*, 151–167. Amsterdam: John Benjamins.
- Bao, Zhiming & Lye Hui Min. 2005. Systemic transfer, topic prominence, and the bare conditional in Singapore English. *Journal of Pidgin and Creole Languages* 20(2). 269–291.
- Bao, Zhiming & Lionel Wee. 1999. The passive in Singapore English. *World Englishes* 18(1). 1–11.
- Barth-Weingarten, Dagmar & Elizabeth Couper Kuhlen. 2011. Action, prosody and emergent constructions: The case of *and*. In Peter Auer & Stefan Pfänder (eds.), *Constructions: Emerging and emergent*, 236–292. Berlin: de Gruyter.

- Bayer, Rachel. 2013. *Null subjects in creole languages*. Ann Arbor, MI: University of Michigan Undergraduate Honors Thesis.
- Bayley, Robert. 2013. Variationist Sociolinguistics. In Robert Bayley, Richard Cameron & Cecilia Lucas (eds.), *The Oxford handbook of Sociolinguistics*, 11–30. Oxford [etc.]: Oxford University Press.
- Bayley, Robert, Norma L. Cardenas, Belinda Trevino Schouten & Carlos M. Velez Salas. 2012. Spanish dialect contact in San Antonio, Texas: An exploratory study. In Kimberly Geeslin & Manuel Diaz-Campos (eds.), *Selected proceedings of the 14th Hispanic Linguistics Symposium*, 48–60. Somerville, MA: Cascadilla Proceedings Project.
- Bayley, Robert, Kristen Greer & Cory Holland. 2013. Lexical frequency and syntactic variation: A test of a linguistic hypothesis. *University of Pennsylvania Working Papers in Linguistics* 19(2). 21–30. <http://repository.upenn.edu/pwpl/vol19/iss2/4>, accessed on 2019-05-06.
- Bayley, Robert & Lucinda Pease-Alvarez. 1997. Null pronoun variation in Mexican descent children's narrative discourse. *Language Variation and Change* 9(3). 349–371.
- Bhatia, Tej K. & William C. Ritchie. 2016. Multilingualism, Hindi-Urdu and Indian English: Intranational and international diaspora. In Gerhard Leitner, Azirah Hashim & Hans-Georg Wolf (eds.), *Communicating with Asia: The future of English as a global language*, 155–169. Cambridge: Cambridge University Press.
- Bhatt, Rakesh M. 2004. Indian English: Syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.), *A handbook of varieties of English*, Volume 2, 1016–1030. Berlin [etc.]: Mouton de Gruyter.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow, England & New York: Pearson Education.
- Biberauer, Theresa (ed.). 2008. *The limits of syntactic variation*. Amsterdam & Philadelphia: John Benjamins.
- Biberauer, Theresa. 2010. Semi null-subject languages, expletives and expletive pro reconsidered. In Theresa Biberauer, Anders Holmberg & Ian Roberts (eds.), *Parametric variation: Null subjects in minimalist theory*, 153–199. Cambridge: Cambridge University Press.
- Biberauer, Theresa, Anders Holmberg & Ian Roberts (eds.). 2010. *Parametric variation: Null subjects in minimalist theory*. Cambridge: Cambridge University Press.
- Bloom, Paul. 1990. Subjectless sentences in child language. *Linguistic Inquiry* 21(4). 491–504.
- Bolker, Ben. 2016. Package 'lme4': Linear Mixed-Effects Models using 'Eigen' and S4 (Version 1.1-12). <http://lme4.r-forge.r-project.org/>, accessed on 2017-08-06.
- Bolton, Kingsley (ed.). 2002. *Hong Kong English: Autonomy and creativity*. Aberdeen & Hong Kong: Hong Kong University Press.
- Bolton, Kingsley. 2003. *Chinese Englishes: A sociolinguistic history*. Cambridge: Cambridge University Press.
- Bolton, Kingsley. 2011. Language policy and planning in Hong Kong: Colonial and post-colonial perspectives. *Applied Linguistics Review* 2. 51–74.
- Bolton, Kingsley & Braj B. Kachru (eds.). 2007. *Asian Englishes*. London & New York: Routledge.
- Breheny, Patrick & Woodrow Burchett. 2017. Package 'visreg': Visualization of regression models (Version 2.4-1). <https://cran.r-project.org/web/packages/visreg/visreg.pdf>, accessed on 2017-08-06.

- 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.
- Brunner, Thomas. 2017. *Simplicity and typological effects in the emergence of New Englishes: The noun phrase in Singaporean and Kenyan English*. Berlin & Boston: de Gruyter.
- Buschfeld, Sarah. 2015. English as a first language in Singapore: A corpus-based analysis of zero subjects. Paper presented at the 2015 Changing Englishes Conference (ChangeE); June 8–10, Helsinki.
- Buschfeld, Sarah. 2018. L1 Singapore English: “Incomplete acquisition”, transmitted L2 system, language change? Paper presented at the 2018 ISLE Conference; July 17–20, London.
- Calhoun, Sasha. 2010. The centrality of metrical structure in signaling information structure. *Language* 86(1). 1–42.
- Camacho, Jose A. 2013. *Null subjects*. Cambridge: Cambridge University Press.
- Cameron, Richard. 1996. A community-based test of a linguistic hypothesis. *Language in Society* 25(1). 61–111.
- Carvalho, Ana M., Rafael Orozco & Naomi L. Shin. 2015. Introduction. In Ana M. Carvalho, Rafael Orozco & Naomi L. Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, xiii–xxvi. Washington, DC: Georgetown University Press.
- Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles N. Li (ed.), *Subject and topic*, 27–55. New York: Academic Press.
- Chafe, Wallace L. 1980. *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.
- Chang, Hui & Zhen Lina. 2018. Asymmetries of null subjects and null objects in L1 English and L1 Japanese learners’ Chinese. *Linguistics* 56(5). 1141–1166.
- Chao, Yuen R. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Chen, Liang & Ning Pan. 2009. Development of English referring expressions in the narratives of Chinese-English bilinguals. *Bilingualism: Language and Cognition* 12(4). 429–445.
- Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.
- Chomsky, Noam. 1982. *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT Press.
- Chu, Chauncey C.-h. 1998. *A discourse grammar of Mandarin Chinese*. New York [etc.]: Lang.
- Clarke, Sandra. 2004. Newfoundland English: Morphology and syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.), *A handbook of varieties of English*, Volume 2, 303–318. Berlin [etc.]: Mouton de Gruyter.
- Cole, Melvyn. 2010. Thematic null subjects and accessibility. *Studia Linguistica* 64(3). 271–320.
- Cote, Sharon A. 1996. *Grammatical and discourse properties of null arguments in English*. Philadelphia, PA: University of Pennsylvania Ph.D. thesis.
- Crewe, William (ed.). 1977. *The English language in Singapore*. Singapore: Eastern Universities Press.
- D’Alessandro, Roberta. 2015. Null subjects. In Antonio Fabregas, Jaume Mateu i Giral & Michael T. Putnam (eds.), *Contemporary linguistic parameters*, 201–226. London & New York: Bloomsbury Academic.
- Davies, Mark. 2013. *Corpus of Global Web-Based English*. <https://www.english-corpora.org/glowbe>, accessed on 2019-06-05.
- Deterding, David. 2007. *Singapore English*. Edinburgh: Edinburgh University Press.

- Deterding, David, Low Ee Ling & Adam Brown (eds.). 2003. *English in Singapore: Research on grammar*. Singapore: McGraw-Hill Education.
- Deterding, David & Gloria Poedjosoedarmo. 1996. To what extent can the ethnic group of young Singaporeans be identified from their speech? In Adam Brown, David Deterding & Low Ee Ling (eds.), *The English language in Singapore: Research on pronunciation*, 1–9. Singapore: SAAL.
- Dryer, Matthew S. 2013. Expression of pronominal subjects. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/101>, accessed on 2019-06-05.
- Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.
- Englebretson, Robert & Marja-Liisa Helasvuo. 2014. Discourse participants in interaction: Cross-linguistic perspectives on subject expression and ellipsis. *Journal of Pragmatics* 63. 1–4.
- Erker, Daniel & Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88(3). 526–557.
- Evans, Stephen. 2009. The evolution of the English-language speech community in Hong Kong. *English World-Wide* 30(3). 278–301.
- Evans, Stephen. 2013. The long march to biliteracy and trilingualism: Language policy in Hong Kong education since the handover. *Annual Review of Applied Linguistics* 33. 302–324.
- Evans, Stephen. 2016. *The English language in Hong Kong: Diachronic and synchronic perspectives*. London: Palgrave Pivot.
- Field, Andy, Jeremy Miles & Zoe Field. 2012. *Discovering statistics using R*. London [etc.]: Sage.
- Filppula, Markku, Juhani Klemola & Devyani Sharma (eds.). 2017. *The Oxford handbook of World Englishes*. Oxford [etc.]: Oxford University Press.
- Flores-Ferran, Nydia. 2007. A bend in the road: Subject personal pronoun expression in Spanish after thirty years of sociolinguistic research. *Language and Linguistics Compass* 1(6). 624–652.
- Foley, Joseph. 1988. Studies in Singapore English: Looking back and looking forward. In Joseph Foley (ed.), *New Englishes: The case of Singapore*, 1–26. Singapore: Singapore University Press.
- Foley, Joseph, Thiru Kandiah, Zhiming Bao, Anthea F. Gupta, Lubna Alsagoff, Chee L. Ho, Lionel Wee, Talib S. Ismail & Wendy D. Bokhorst-Heng (eds.). 1998. *English in new cultural contexts: Reflections from Singapore*. Singapore [etc.]: Oxford University Press.
- Fox, John. 2016a. Package ‘car’: Companion to Applied Regression (Version 2.1-4). <https://r-forge.r-project.org/projects/car/>, accessed on 2017-08-06.
- Fox, John. 2016b. Package ‘effects’: Effect displays for linear, generalized linear, and other models (Version 3.1-2). <https://cran.r-project.org/web/packages/effects/effects.pdf>, accessed on 2017-08-06.
- Frascarelli, Mara. 2007. Subjects, topics and the interpretation of referential pro: An interface approach to the linking of (null) pronouns. *Natural Language and Linguistic Theory* 25(4). 691–734.
- Fuller, Judith W. & Jeanette K. Gundel. 1987. Topic-prominence in interlanguage. *Language Learning* 37(1). 1–18.

- Gargesh, Ravinder. 2004. Indian English: Phonology. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.), *A handbook of varieties of English*, Volume 1, 992–1002. Berlin [etc.]: Mouton de Gruyter.
- Gargesh, Ravinder. 2006. South Asian Englishes. In Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds.), *A handbook of World Englishes*, 90–113. Malden, MA [etc.]: Blackwell.
- Gargesh, Ravinder & Sailaja Pingali. 2017. South Asia. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*, 425–446. Oxford [etc.]: Oxford University Press.
- Gil, David. 2001. Escaping Eurocentrism: Fieldwork as a process of unlearning. In Paul Newman & Martha Ratliff (eds.), *Linguistic fieldwork*, 102–132. Cambridge: Cambridge University Press.
- Gilligan, Gary. 1987. *A cross-linguistic approach to the pro-drop parameter*. Los Angeles, CA: University of Southern California Ph.D. Dissertation.
- Gisborne, Nikolas. 2000. Relative clauses in Hong Kong English. *World Englishes* 19(3). 357–371.
- Gisborne, Nikolas. 2009. Aspects of the morphosyntactic typology of Hong Kong English. *English World-Wide* 30(2). 149–169.
- Givon, Talmy. 1979a. From discourse to syntax: Grammar as a processing strategy. In Talmy Givon (ed.), *Discourse and syntax*, 81–112. New York: Academic Press.
- Givon, Talmy. 1979b. *On understanding grammar*. New York: Academic Press.
- Givon, Talmy. 1983. Topic continuity in discourse: An introduction. In Talmy Givon (ed.), *Topic continuity in discourse: A quantitative cross-language study*, 1–41. Amsterdam: Benjamins.
- Givon, Talmy. 1995. Isomorphism in the grammatical code: Cognitive and biological considerations. In Raffaele Simone (ed.), *Iconicity in Language*, 47–76. Amsterdam: John Benjamins Publishing Company.
- Green, Eugene & Charles F. Meyer (eds.). 2014. *The variability of current World Englishes*. Berlin: De Gruyter Mouton.
- Greenbaum, Sidney. 1996. Introducing ICE. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 3–12. Oxford: Clarendon Press.
- Gries, Stefan Th. 2005. Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34(4). 365–399.
- Gries, Stefan Th. 2010. Collostructional analysis: Computing the degree of association between words and words/constructions [Coll.analysis] (Version 3.2a). <http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/coll.analysis.r>, accessed on 2017-08-07.
- Gries, Stefan Th. 2013. *Statistics for linguistics with R: A practical introduction*, 2nd edn. Berlin & Boston: Walter de Gruyter.
- Gundel, Jeanette K., Mamadou Bassene, Bryan Gordon, Linda Humnick & Amel Khalfaoui. 2010. Testing predictions of the Givenness Hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics* 42(7). 1770–1785.
- Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.
- Gupta, Anthea F. 1991. Almost a creole: Singapore Colloquial English. *California Linguistic Notes* 23. 9–21.
- Gupta, Anthea F. 1994. *The step-tongue: Childrens' English in Singapore*. Clevedon: Multilingual Matters.

- Gupta, Anthea F. 1997. Colonisation, migration, and functions of English. In Edgar W. Schneider (ed.), *Englishes around the world*, 47–58. Amsterdam: John Benjamins.
- Gupta, Anthea F. 1998. The situation of English in Singapore. In Joseph Foley, Thiru Kandiah, Zhiming Bao, Anthea F. Gupta, Lubna Alsagoff, Chee L. Ho, Lionel Wee, Talib S. Ismail & Wendy D. Bokhorst-Heng (eds.), *English in new cultural contexts: Reflections from Singapore*, 106–126. Singapore [etc.]: Oxford University Press.
- Haegeman, Liliane & Tabea Ihsane. 1999. Subject ellipsis in embedded clauses in English. *English Language and Linguistics* 3(1). 117–145.
- Haegeman, Liliane & Tabea Ihsane. 2001. Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language Acquisition* 9(4). 329–346.
- Halliday, Michael A. K. 1967. Note on transitivity and theme in English: Part 2. *Journal of Linguistics* 2(3). 199–244.
- Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
- Hansen, Beke. 2018. *Corpus linguistics and sociolinguistics: A study of variation and change in the modal systems of World Englishes*. Leiden & Boston: Brill Rodopi.
- Harrel, Frank E. Jr. Package ‘rms’: Regression Modeling Strategies (Version 5.0-0). <http://cran.r-project.org/web/packages/rms/>, accessed on 2017-08-06.
- Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers Linguistiques d’Ottawa / Ottawa Papers in Linguistics* 26. 15–25. <http://artsites.uottawa.ca/cloopl/doc/Null-Subject-in-English-Wonder-if-it-exists.pdf>, accessed on 2017-09-09.
- Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook = Sprachtypologie und sprachliche Universalien*, 1492–1510. Berlin: de Gruyter.
- Haspelmath, Martin & the APICS Consortium. 2013a. Expletive subject of existential verb. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apicsonline.info/parameters/64>, accessed on 2017-09-09.
- Haspelmath, Martin & the APICS Consortium. 2013b. Expression of pronominal subjects. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://apics-online.info/parameters/62#2/30.4/10.0>, accessed on 2019-05-06.
- Heine, Bernd & Tania Kuteva. 2005. *Language contact and grammatical change*. Cambridge: Cambridge University Press.
- Helasvuoto, Marja-Liisa & Aki-Juhani Kyrolainen. 2016. Choosing between zero and pronominal subject: Modeling subject expression in the 1st person singular in Finnish conversation. *Corpus Linguistics and Linguistic Theory* 12(2). 263–300.
- Heusinger, Klaus von. 1999. *Intonation and information structure*. Konstanz: University of Konstanz Postdoctoral Thesis.
- Hickey, Raymond. 2010. Contact and language shift. In Raymond Hickey (ed.), *The handbook of language contact*, 149–169. Malden, MA: Wiley-Blackwell.
- Hinterholz, Roland & Svetlana Petrova (eds.). 2009. *Information structure and language change: New approaches to word order variation in Germanic*. Berlin: Mouton de Gruyter.
- Hirschberg, Julia. 2004. Pragmatics and intonation. In Laurence R. Horn & Gregory Ward (eds.), *The handbook of pragmatics*, 515–537. Malden, MA: Blackwell.

- Ho, Chee Lick & Lubna Alsagoff. 1998. English as the common language in multicultural Singapore. In Joseph Foley, Thiru Kandiah, Zhiming Bao, Anthea F. Gupta, Lubna Alsagoff, Chee L. Ho, Lionel Wee, Talib S. Ismail & Wendy D. Bokhorst-Heng (eds.), *English in new cultural contexts: Reflections from Singapore*, 201–207. Singapore [etc.]: Oxford University Press.
- Ho, Mian-Lian & John T. Platt. 1993. *Dynamics of a contact continuum: Singaporean English*. Oxford: Clarendon.
- Holmberg, Anders. 2010. Null subject parameters. In Theresa Biberauer, Anders Holmberg & Ian Roberts (eds.), *Parametric variation: Null subjects in minimalist theory*, 88–124. Cambridge: Cambridge University Press.
- Holmberg, Anders, Aarti Nayudu & Michelle Sheehan. 2009. Three partial null-subject languages: A comparison of Brazilian Portuguese, Finnish and Marathi. *Studia Linguistica* 63(1). 59–97.
- Hopper, Paul. 2002. Hendiadys and auxiliation in English. In Joan L. Bybee & Michael Noonan (eds.), *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson*, 145–174. Amsterdam: John Benjamins.
- Huang, C. T. J. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15(4). 531–574.
- Huang, Yan. 1994. *The syntax and pragmatics of anaphora: A study with special reference to Chinese*. Cambridge: Cambridge University Press.
- Huang, Yan. 2000. *Anaphora: A cross-linguistic study*. Oxford & New York: Oxford University Press.
- Huang, Yan. 2014. *Pragmatics*. Oxford: Oxford University Press.
- Huddleston, Rodney D., Geoffrey K. Pullum & Laurie Bauer (eds.). 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
- Hung, Tony T. 2012. Hong Kong English. In Low Ee Ling & Azirah Hashim (eds.), *English in Southeast Asia: Features, policy and language in use*, 113–133. Amsterdam: Benjamins.
- Hyams, Nina. 1986. *Language acquisition and the theory of parameters*. Dordrecht [etc.]: Reidel.
- Hyams, Nina. 1994. V2 null arguments and COMP projections. In Teun Hoekstra & Bonnie D. Schwartz (eds.), *Language acquisition studies in generative grammar: 14th GLOW conference papers*, 21–55. Amsterdam & Philadelphia: John Benjamins.
- Hyams, Nina & Kenneth Wexler. 1993. On the grammatical basis of null subjects in child language. *Linguistic Inquiry* 24(3). 421–459.
- Hyland, Ken. 1997. Language attitudes at the handover: Communication and identity in 1997 Hong Kong. *English World-Wide* 18(2). 191–210.
- Ihalainen, Ossi. 1991. The grammatical subject in educated and dialectal English: Comparing the London-Lund Corpus and the Helsinki Corpus of Modern English dialects. In Stig Johansson & Anna-Brita Stenstrom (eds.), *English computer corpora: Selected papers and research guide*, 201–214. Berlin & New York: Mouton de Gruyter.
- Jaeggli, Osvaldo & Kenneth J. Safir. 1989a. The null subject parameter and parametric theory. In Osvaldo Jaeggli & Kenneth J. Safir (eds.), *The null subject parameter*, 1–44. Dordrecht [etc.]: Kluwer.
- Jaeggli, Osvaldo & Kenneth J. Safir (eds.). 1989b. *The null subject parameter*. Dordrecht [etc.]: Kluwer.

- Jia, Li & Robert Bayley. 2002. Null pronoun variation in Mandarin Chinese. *University of Pennsylvania Working Papers in Linguistics* 8(3). <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1748&context=pwpl>, accessed on 2019-05-06.
- Jin, Hong G. 1994. Topic-prominence and subject-prominence in L2 acquisition: Evidence of English-to-Chinese typological transfer. *Language Learning* 44(1). 101–122.
- Johnson, Daniel Ezra. 2009. Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass* 3(1). 359–383.
- Johnson, Daniel Ezra. 2016. Rbrul (Version 2.3.1). <http://www.danielezrajohnson.com/rbrul.html>, accessed on 2017-08-07.
- Junghare, Indira Y. 1988. Topic-prominence and zero NP-anaphora in Marathi and Hindi. In Mohammad A. Jazayery & Werner Winter (eds.), *Languages and cultures*, 309–328. Berlin [etc.]: Mouton de Gruyter.
- Kachru, Braj B. 1965. The Indianness in Indian English. *Word* 21(3). 391–400. Kachru, Braj B. (ed.). 1982. *The other tongue: English across cultures*. Urbana [etc.]: University of Illinois Press.
- Kachru, Braj B. 1983. *The Indianization of English: The English language in India*. Delhi: Oxford University Press.
- Kachru, Braj B. 1985. Standards, codification, and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk & Henry G. Widdowson (eds.), *English in the world: Teaching and learning the language and the literature*, 11–30. Cambridge: Cambridge University Press.
- Kachru, Braj B. 1986. *The alchemy of English: The spread, functions and models of non-native Englishes*. Oxford [etc.]: Pergamon Institute of English.
- Kachru, Braj B. 2005. *Asian Englishes: Beyond the canon*. Hong Kong: Hong Kong University Press.
- Kachru, Braj B., Yamuna Kachru & Cecil L. Nelson (eds.). 2006. *A handbook of World Englishes*. Malden, MA [etc.]: Blackwell.
- Kachru, Yamuna & Cecil L. Nelson. 2006. *World Englishes in Asian contexts*. Hong Kong: Hong Kong University Press.
- Keenan, Edward L. 1976. Towards a universal definition of “subject”. In Charles N. Li (ed.), *Subject and topic*, 303–334. New York: Academic Press.
- Kibrik, Aleksandr E. 2001. Subject-oriented vs. subjectless languages. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook = Sprachtypologie und sprachliche Universalien*, 1413–1423. Berlin: de Gruyter.
- Kim, Chonghyuck & Yosuke Sato. 2013. More on *kena*-passives in Singapore English. *World Englishes* 32(3). 297–307.
- Kirkpatrick, Andy & Roland Sussex (eds.). 2012. *English as an international language in Asia: Implications for language education*. Dordrecht: Springer Netherlands.
- Kiss, Katalin E. 2001. Discourse configurationality. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook = Sprachtypologie und sprachliche Universalien*, 1442–1455. Berlin: de Gruyter.
- Kortmann, Bernd. 2014. Comparison and comparability: The WAVE perspective. In Silvia Mergenthal & Reingard M. Nischik (eds.), *Anglistentag 2013 Konstanz: Proceedings*, 323–336. Trier: Wissenschaftlicher Verlag.

- Kortmann, Bernd & Kerstin Lunkenheimer (eds.). 2012. *The Mouton world atlas of variation in English*. Berlin: De Gruyter Mouton.
- Kortmann, Bernd & 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 2019-05-06.
- Kortmann, Bernd, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.). 2004. *A handbook of varieties of English*. Volume 1: Phonology. Volume 2: Morphology and Syntax. Berlin [etc.]: Mouton de Gruyter.
- Kortmann, Bernd & Verena Schröter. 2017. Varieties of English. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 304–330. Cambridge: Cambridge University Press.
- Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: Morphological and syntactic variation in English. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.), *A handbook of varieties of English*, Volume 2, 1142–1202. Berlin [etc.]: Mouton de Gruyter.
- Kortmann, Bernd & Benedikt Szmrecsanyi. 2009. World Englishes between simplification and complexification. In Thomas Hoffmann & Lucia Siebers (eds.), *World Englishes: Problems, properties and prospects*, 263–286. Amsterdam: John Benjamins.
- Kortmann, Bernd & Christoph Wolk. 2012. Morphosyntactic variation in the anglophone world: A global perspective. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 906–936. Berlin: De Gruyter Mouton.
- Krishnamurti, Bhadriraju. 2003. *The Dravidian languages*. Cambridge [etc.]: Cambridge University Press.
- Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4). 715–762.
- Labov, William. 1982. Building on empirical foundations. In Winfred Lehmann & Yakov Malkiel (eds.), *Perspectives on historical linguistics*, 17–92. Amsterdam & Philadelphia: John Benjamins.
- Lambrecht, Knud. 1988. Presentational cleft constructions in spoken French. In John Haiman & Sandra A. Thompson (eds.), *Clause combining in grammar and discourse*, 135–180. Amsterdam: John Benjamins Publishing Company.
- Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.
- Lange, Claudia. 2012. *The syntax of spoken Indian English*. Amsterdam: Benjamins.
- Lee, Nala H. 2014. *A grammar of Baba Malay with sociophonetic considerations*. Manoa: University of Hawai'i Ph.D. thesis.
- Lee, Nala H., Ling A. Ping & Hiroki Nomoto. 2009. Colloquial Singapore English *got*: Functions and substratal influences. *World Englishes* 28(3). 293–318.
- Leimgruber, Jakob R. E. 2009. *Modelling variation in Singapore English*. Oxford: University of Oxford Ph.D. thesis.
- Leimgruber, Jakob R. E. 2013. *Singapore English: Structure, variation, and usage*. Cambridge: Cambridge University Press.
- Leong Ping, Alvin. 2003. Subject omission in Singapore Colloquial English. In David Deterding, Low Ee Ling & Adam Brown (eds.), *English in Singapore: Research on grammar*, 11–21. Singapore: McGraw-Hill Education.

- Leroux, Martine & Lidia-Gabriela Jarmasz. 2006. A study about nothing: Null subjects as a diagnostic of convergence between English and French. *University of Pennsylvania Working Papers in Linguistics* 12(2). <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1669&context=pwpl>, accessed on 2019-05-06.
- Levshina, Natalia. 2015a. *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam & Philadelphia: John Benjamins.
- Levshina, Natalia. 2015b. Package 'RLing' (Version 1.0). <https://benjamins.com/sites/z.195/content/package.html>, accessed on 2017-08-07.
- Lezama, Carlos G. & Amit Almor. 2011. Repeated names, overt pronouns, and null pronouns in Spanish. *Language and Cognitive Processes* 26(3). 437–454.
- Li, Charles N. & Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li (ed.), *Subject and topic*, 457–490. New York: Academic Press.
- Li, Charles N. & Sandra A. Thompson. 1979. Third-person pronouns and zero anaphora in Chinese discourse. In Talmy Givón (ed.), *Discourse and syntax*, 311–335. New York: Academic Press.
- Li, Charles N. & Sandra A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley & Los Angeles: University of California Press.
- Li, Michelle & Stephen Matthews. 2013. Chinese Pidgin English structure dataset. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apics-online.info/contributions/20>, accessed on 2019-05-06.
- Li, Xiaoshi. 2014. Variation in subject pronominal expression in L2 Chinese. *Studies in Second Language Acquisition* 36(1). 39–68.
- Li, Xiaoshi, Xiaoqing Chen & Wen-Hsing Chen. 2012. Variation of subject pronominal expression in Mandarin Chinese. *Sociolinguistic Studies* 6(1). 91–119.
- Liceras, Juana M. & Lourdes Díaz. 1999. Topic-drop versus pro-drop: Null subjects and pronominal subjects in the Spanish L2 of Chinese, English, French, German and Japanese speakers. *Second Language Research* 15(1). 1–40.
- Lim, Lisa. 1996. Ethnic group differences aligned? Intonation patterns of Chinese, Indian and Malay Singaporean English. In Adam Brown, David Deterding & Low Ee Ling (eds.), *The English language in Singapore: Research on pronunciation*, 10–21. Singapore: SAAL.
- Lim, Lisa. 2001. Ethnic groups varieties of Singapore English: Melody or harmony? In Vincent B. Y. Ooi (ed.), *Evolving identities: The English language in Singapore and Malaysia*, 53–68. Singapore: Times Academic Press.
- Lim, Lisa (ed.). 2004. *Singapore English: A grammatical description*. Amsterdam: Benjamins.
- Lim, Lisa. 2007. Mergers and acquisitions: On the ages and origins of Singapore English particles. *World Englishes* 26(4). 446–473.
- Lim, Lisa. 2009. Not just an “Outer Circle”, “Asian” English: Singapore English and the significance of ecology. In Thomas Hoffmann & Lucia Siebers (eds.), *World Englishes: Problems, properties and prospects*, 179–206. Amsterdam: John Benjamins.
- Lim, Lisa. 2011. Tone in Singlish. In Claire Lefebvre (ed.), *Creoles, their substrates, and language typology*, 271–287. Amsterdam: John Benjamins.
- Lim, Lisa & Umberto Ansaldi. 2016. *Languages in contact*. Cambridge: Cambridge University Press.
- Lim, Lisa & Joseph Foley. 2004. English in Singapore and Singapore English: Background and methodology. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 1–18. Amsterdam: Benjamins.

- Lim, Lisa & Nikolas Gisborne (eds.). 2011. *The typology of Asian Englishes*. Amsterdam & Philadelphia: John Benjamins.
- Lisser, Tamirand N. de, Stephanie Durrleman, Luigi Rizzi & Ur Shlonsky. 2015. The acquisition of Jamaican Creole: Null subject phenomenon. *Language Acquisition* 23(3). 261–292.
- Liu, Dilin. 2008. Adequate language description in L2 research/teaching: The case of pro-drop language speakers learning L2 English. *International Journal of Applied Linguistics* 18(3). 274–292.
- Loar, Jian K. 2011. *Chinese syntactic grammar: Functional and conceptual principles*. New York: Peter Lang.
- Low, Ee Ling & Azirah Hashim (eds.). 2012. *English in Southeast Asia: Features, policy and language in use*. Amsterdam: Benjamins.
- Luke, Kang-kwong & Jack C. Richards. 1982. English in Hong Kong: Functions and status. *English World-Wide* 3(1). 47–64.
- Lunkenheimer, Kerstin. 2012. Typological profile: L2 varieties. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 844–873. Berlin: De Gruyter Mouton.
- Mair, Christian. 2003. Kreolismen und verbales Identitätsmanagement im geschriebenen jamaikanischen Englisch. In Elisabeth Vogel, Antonia Napp & Wolfram Lutterer (eds.), *Zwischen Ausgrenzung und Hybridisierung*, 79–96. Würzburg: Ergon.
- Mair, Christian. 2017. World Englishes and corpora. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*, 103–120. Oxford [etc.]: Oxford University Press.
- Matras, Yaron. 2009. *Language contact*. Cambridge: Cambridge University Press.
- Matthews, Stephen. 2010. Language contact and Chinese. In Raymond Hickey (ed.), *The handbook of language contact*, 757–769. Malden, MA: Wiley-Blackwell.
- Matthews, Stephen & Virginia Yip. 1994. *Cantonese: A comprehensive grammar*. London & New York: Routledge.
- Mereu, Lunella (ed.). 2009. *Information Structure and its interfaces*. Berlin: Mouton de Gruyter.
- Mesthrie, Rajend. 1992. *English in language shift: The history, structure and sociolinguistics of South African Indian English*. Cambridge [etc.]: Cambridge University Press.
- Mesthrie, Rajend. 2006. Anti-deletions in a second language variety: A study of Black South African English mesolect. *English World-Wide* 27(2). 111–146.
- Mesthrie, Rajend. 2012. Regional profile: Asia. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 784–805. Berlin: De Gruyter Mouton.
- Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties*. Cambridge [etc.]: Cambridge University Press.
- Meyerhoff, Miriam. 2000. The emergence of creole subject-verb agreement and the licensing of null subjects. *Language Variation and Change* 12(2). 203–230.
- Meyerhoff, Miriam. 2009. Replication, transfer, and calquing: Using variation as a tool in the study of language contact. *Language Variation and Change* 21(3). 297–317.
- Meyerhoff, Miriam. 2013. Bislama structure dataset. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apics-online.info/contributions/23>, accessed on 2019-05-06.
- Michaelis, Susanne M. & the APiCS Consortium. 2013. Expletive subject in ‘seem’ constructions. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath &

- Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://apics-online.info/parameters/63#2/30.3/5.0>, accessed on 2019-05-06.
- Michaelis, Susanne M., Philippe Maurer, Martin Haspelmath & 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 2019-05-06.
- Moag, Rodney F. & Robert Poletto. 1991. Discourse level evidence for south Asia as a linguistic area. *Language Sciences* 13(2). 229–254.
- Modesto, Marcello. 2008. Topic prominence and null subjects. In Theresa Biberauer (ed.), *The limits of syntactic variation*, 375–410. Amsterdam & Philadelphia: John Benjamins.
- Mufwene, Salikoko S. 2008. *Language evolution: Contact, competition and change*. London & New York: Continuum.
- Mufwene, Salikoko S. 2001. *The ecology of language evolution*. Cambridge: Cambridge University Press.
- Mukherjee, Joybrato. 2007. Steady states in the evolution of New Englishes: Present day Indian English as an equilibrium. *Journal of English Linguistics* 35(2). 157–187.
- Mukherjee, Joybrato & Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1). 27–51.
- Mukherjee, Joybrato & Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27(2). 147–173.
- Murata, Kumiko & Jennifer Jenkins (eds.). 2009. *Global Englishes in Asian contexts: Current and future debates*. Basingstoke & New York: Palgrave Macmillan.
- Nagy, Naomi. 2015. A sociolinguistic view of null subjects and VOT in Toronto heritage languages. *Lingua* 164. 309–327.
- Nagy, Naomi G., Nina Aghdasi, Derek Denis & Alexandra Motut. 2011. Null subjects in heritage languages: Contact effects in a cross-linguistic context. *University of Pennsylvania Working Papers in Linguistics* 17(2). Article 16. <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1202&context=pwpl>, accessed on 2019-05-06.
- Nariyama, Shigeko. 2004. Subject ellipsis in English. *Journal of Pragmatics* 36(2). 237–264.
- Neeleman, Ad & Kriszta Szendroi. 2007. Radical pro drop and the morphology of pronouns. *Linguistic Inquiry* 38(4). 671–714.
- Nelson, Gerald. 1996. The design of the corpus. In Sidney Greenbaum (ed.), *Comparing English worldwide: The International Corpus of English*, 27–35. Oxford: Clarendon Press.
- Nelson, Gerald. 2006. *The ICE Hong Kong corpus: User manual*. Distributed with the International Corpus of English, ICE-Hong Kong.
- Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam & Philadelphia: John Benjamins Publishing Company.
- Newmeyer, Frederick J. 2005. *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.
- Nicolis, Marco. 2008. The null subject parameter and correlating properties: The case of Creole languages. In Theresa Biberauer (ed.), *The limits of syntactic variation*, 271–294. Amsterdam & Philadelphia: John Benjamins.
- Oh, Sun-Young. 2005. English zero anaphora as an interactional resource. *Research on Language & Social Interaction* 38(3). 267–302.

- Oh, Sun-Young. 2006. English zero anaphora as an interactional resource II. *Discourse Studies* 8(6). 817–846.
- Ooi, Vincent B. Y. 2001. Upholding standards or passively observing language? Corpus evidence and the Concentric Circles Model. In Vincent B. Y. Ooi (ed.), *Evolving identities: The English language in Singapore and Malaysia*, 168–183. Singapore: Times Academic Press.
- Otheguy, Ricardo & Ana C. Zentella. 2012. *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. Oxford, New York: Oxford University Press.
- Otheguy, Ricardo, Ana C. Zentella & David Livert. 2007. Language and dialect contact in Spanish in New York: Towards the formation of a speech community. *Language* 83(1). 1–33.
- Otheguy, Ricardo, Ana C. Zentella & David Livert. 2010. Generational differences in pronominal usage in Spanish reflecting language and dialect contact in a bilingual setting. In Muriel Norde, Bob d. Jonge & Cornelius Hasselblatt (eds.), *Language contact: New perspectives*, 45–62. Amsterdam: John Benjamins.
- Pakir, Anne. 1991. The range and depth of English-knowing bilinguals in Singapore. *World Englishes* 10(2). 167–179.
- Pang, Terrence. 2003. Hong Kong English: A stillborn variety? *English Today* 19(2). 12–18.
- Parviainen, Hanna. 2015. Subject omission in World Englishes. Paper presented at the 2015 Changing Englishes Conference (ChangE); June 8–10, Helsinki.
- Paul, Premila. 2003. The Master's language and its Indian uses. In Christian Mair (ed.), *The politics of English as a world language: New horizons in postcolonial cultural studies*, 359–366. Amsterdam & New York: Rodopi.
- Percillier, Michael. 2016. *World Englishes and second language acquisition: Insights from Southeast Asian Englishes*. Amsterdam & Philadelphia: John Benjamins.
- Perlmutter, David. 1971. *Deep and surface constraints in generative grammar*. New York: Holt, Rinehart and Winston.
- Phinney, Marianne. 1987. The pro-drop parameter in second language acquisition. In Thomas Roeper, Kenneth Wexler & Edwin Williams (eds.), *Parameter setting*, 221–238. Dordrecht: Springer Netherlands.
- Pingali, Sailaja. 2009. *Indian English*. Edinburgh: Edinburgh University Press.
- Platt, John T. 1975. The Singapore English speech continuum and its basilect 'Singlish' as a creoloid. *Anthropological Linguistics* 17. 363–374.
- Platt, John T. 1979. Variation and implicational relationships: Copula realization in Singapore. *General Linguistics* 19. 1–14.
- Platt, John T. & Mian-Lian Ho. 1989. Discourse particles in Singaporean English: Substratum influences and universals. *World Englishes* 8(2). 215–221.
- Platt, John T. & Heidi Weber. 1980. *English in Singapore and Malaysia: Status, features, functions*. Oxford [etc.]: Oxford University Press.
- Platt, John T., Heidi Weber & Mian-Lian Ho. 1983. *Singapore and Malaysia*. Amsterdam: Benjamins.
- Platt, John T., Heidi Weber & Mian-Lian Ho. 1984. *The New Englishes*. London: Routledge and Kegan Paul.
- Poplack, Shana. 1980. Deletion and disambiguation in Puerto Rican Spanish: Competing constraints on (s) deletion. *Language* 56(2). 371–385.
- Poplack, Shana & Marjory Meechan. 1998. How languages fit together in code-mixing. *International Journal of Bilingualism* 2(2). 127–138.
- Poplack, Shana & Sali Tagliamonte. 1993. The zero-marked verb: Testing the creole hypothesis. *Journal of Pidgin and Creole Languages* 8(2). 171–206.

- Poplack, Shana & Sali Tagliamonte. 2001. *African American English in the diaspora*. Oxford: Basil Blackwell.
- Prescott, David L., Isabel P. Martin, Andy Kirkpatrick & Azirah Hashim (eds.). 2007. *English in Southeast Asia: Varieties, literacies and literatures*. Newcastle upon Tyne: Cambridge scholars.
- Prince, Ellen F. 1992. The ZPG letter: Subjects, definiteness, and information-status. In Sandra A. Thompson & William C. Mann (eds.), *Discourse description: Diverse analyses of a fundraising text*, 295–325. Amsterdam & Philadelphia: John Benjamins.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1972. *A grammar of contemporary English*. London: Longman.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London & New York: Longman.
- Register, Norma. 1990. Influences of typological parameters on L2 learners' judgements of null pronouns in English. *Language Learning* 40(3). 369–383.
- Richards, Jack C. 1977. Variation in Singapore English. In William Crewe (ed.), *The English language in Singapore*, 68–82. Singapore: Eastern Universities Press.
- Rizzi, Luigi. 1982. *Issues in Italian Syntax*. Dordrecht: Foris Publications.
- Rizzi, Luigi. 1986. Null objects in Italian and the theory of pro. *Linguistic Inquiry* 17(3). 501–557.
- Rizzi, Luigi. 1994. Early null subjects and root null subjects. In Teun Hoekstra & Bonnie D. Schwartz (eds.), *Language acquisition studies in generative grammar: 14th GLOW conference papers*, 151–176. Amsterdam & Philadelphia: John Benjamins.
- Rizzi, Luigi. 2000a. *Comparative syntax and language acquisition*. London: Routledge.
- Rizzi, Luigi. 2000b. Remarks on early null subjects. In Marc-Ariel Friedeman & Luigi Rizzi (eds.), *The acquisition of syntax: Studies in comparative developmental linguistics*, 269–292. New York: Longman.
- Roberts, Ian & Anders Holmberg. 2010. Introduction: Parameters in minimalist theory. In Theresa Biberauer, Anders Holmberg & Ian Roberts (eds.), *Parametric variation: Null subjects in minimalist theory*, 1–57. Cambridge: Cambridge University Press.
- Rubdy, Rani. 2001. Creative destruction: Singapore's Speak Good English movement. *World Englishes* 20(3). 341–355.
- Sankoff, David. 1988. Sociolinguistics and syntactic variation. In Frederick J. Newmeyer (ed.), *Linguistics: The Cambridge survey*, 140–161. New York: Cambridge University Press.
- Sankoff, David. 2008. Variable rules. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier & Peter Trudgill (eds.), *Sociolinguistics: An international handbook of the science of language and society = Soziolinguistik: ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, 2nd edn. Berlin: De Gruyter Reference Global.
- Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2). 189–222.
- Sato, Yosuke. 2013. Wh-Questions in Colloquial Singapore English: Adaptive traits from vernacular Malay and typological congruence. *Journal of Pidgin and Creole Languages* 28(2). 299–322.
- Sato, Yosuke. 2014. Argument ellipsis in Colloquial Singapore English and the Anti-Agreement Hypothesis. *Journal of Linguistics* 50(2). 365–401.
- Sato, Yosuke & Chonghyuck Kim. 2012. Radical pro drop and the role of syntactic agreement in Colloquial Singapore English. *Lingua* 122. 858–873.
- Schilk, Marco. 2011. *Structural nativization in Indian English lexicogrammar*. Amsterdam: Benjamins.

- Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin & New York: Mouton de Gruyter.
- Schneider, Edgar W. 2007. *Postcolonial Englishes: Varieties around the world*. Cambridge [etc.]: Cambridge University Press.
- Schneider, Edgar W. 2017. Models of English in the world. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*, 35–60. Oxford [etc.]: Oxford University Press.
- Schröter, Verena. 2010. *Stylistic variation in Singapore English: A corpus-based analysis of deletion and subject omission*. Freiburg: Albert Ludwigs Universität MA thesis.
- Schröter, Verena. 2012. Colloquial Singaporean English (Singlish). In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 562–572. Berlin: De Gruyter Mouton.
- Schröter, Verena & Bernd Kortmann. 2016. Pronoun deletion in Hong Kong English and Colloquial Singaporean English. *World Englishes* 35(2). 221–241.
- Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and change*. Amsterdam: John Benjamins.
- Setter, Jane, Brian H.-S. Chan & Cathy S. P. Wong. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.
- Sharma, Devyani. 2005. Language transfer and discourse universals in Indian English article use. *Studies in Second Language Acquisition* 27(4). 535–566.
- Sharma, Devyani. 2009. Typological diversity in New Englishes. *English World-Wide* 30(2). 170–195.
- Sharma, Devyani. 2012a. Indian English. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 523–530. Berlin: De Gruyter Mouton.
- Sharma, Devyani. 2012b. Shared features in New Englishes. In Raymond Hickey (ed.), *Areal features of the anglophone world*, 211–232. Berlin & New York: Mouton.
- Sharma, Devyani & John R. Rickford. 2009. AAVE/creole copula absence: A critique of the imperfect learning hypothesis. *Journal of Pidgin and Creole Languages* 24(1). 53–90.
- Shastri, S. V. 2002. *Overview of the Indian Component of the International Corpus of English (ICE-India)*. Distributed with the Indian section of the International Corpus of English, ICE-India.
- Siegel, Jeff. 2006. Links between SLA and creole studies: Past and present. In Claire Lefebvre, Lydia White & Christine Jourdan (eds.), *L2 acquisition and creole genesis: Dialogues*, 15–46. Amsterdam [etc.]: John Benjamins.
- Siegel, Jeff. 2008. *The emergence of pidgin and creole languages*. Oxford & New York: Oxford University Press.
- Siemund, Peter. 2016. The mutual relevance of typology and variation studies. *Linguistic Typology* 20(3). 521–536.
- Siemund, Peter. 2013. *Varieties of English: A typological approach*. Cambridge: Cambridge University Press.
- Siemund, Peter & Julia Davydova. 2017. World Englishes and the study of typology and universals. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*, 123–146. Oxford [etc.]: Oxford University Press.
- Siemund, Peter, Monika E. Schulz & Martin Schweinberger. 2014. Studying the linguistic ecology of Singapore: A comparison of college and university students. *World Englishes* 33(3). 340–362.
- Sigurdsson, Halldor A. 2011. Conditions on argument drop. *Linguistic Inquiry* 42(2). 267–304.

- Silva-Corvalan, Carmen. 1994. *Language contact and change: Spanish in Los Angeles*. New York: Oxford University Press.
- Spring, Ryan & Kaoru Horie. 2013. How cognitive typology affects second language acquisition: A study of Japanese and Chinese learners of English. *Cognitive Linguistics* 24(4). 689–710.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43.
- Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1). 113–150.
- Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin & New York: Mouton de Gruyter.
- Szmrecsanyi, Benedikt & Bernd Kortmann. 2009. Vernacular universals and angloversals in a typological perspective. In Markku Filpula, Juhani Klemola & Heli Paulasto (eds.), *Vernacular universals and language contacts: Evidence from varieties of English and beyond*, 33–55. New York [etc.]: Routledge.
- Taboada, Maite & Loreley Wiesemann. 2010. Subjects and topics in conversation. *Journal of Pragmatics* 42(7). 1816–1828.
- Tagliamonte, Sali. 2006. *Analysing sociolinguistic variation*. Cambridge [etc.]: Cambridge University Press.
- Tagliamonte, Sali. 2008. Comparative sociolinguistics. In J. K. Chambers & Natalie Schilling-Estes (eds.), *The handbook of language variation and change*, 128–156. Chichester, West Sussex, UK [etc.]: Wiley-Blackwell.
- Tagliamonte, Sali & 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). 135–178.
- Tamaredo, Ivan & Teresa Fanego. 2016. Pronoun omission and agreement: An analysis based on ICE Singapore and ICE India. *ICAME Journal* 40. 95–118.
- Tan, Ludwig. 2003. Topic prominence and null arguments in Singapore Colloquial English. In David Deterding, Low Ee Ling & Adam Brown (eds.), *English in Singapore: Research on grammar*, 1–10. Singapore: McGraw-Hill Education.
- Tao, Liang & Alice F. Healy. 2005. Zero anaphora: Transfer of reference tracking strategies from Chinese to English. *Journal of Psycholinguistic Research* 34(2). 99–131.
- Taraldsen, Knut T. 1978. The scope of *wh*-movement in Norwegian. *Linguistic Inquiry* 9(4). 623–640.
- Teddman, Laura. 2011. Subject ellipsis by text type: An investigation using ICE-GB. In John Newman, R. Harald Baayen & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 71–88. Amsterdam & New York: Rodopi.
- Teddman, Laura & John Newman. 2007. Subject ellipsis in English: Construction of and findings from a diary corpus. In Catherine Camugli, Matthieu Constant & Anne Dister (eds.), *Proceedings from the 26th conference on Lexis and Grammar*. <https://www.semanticscholar.org/paper/Subject-Ellipsis-in-English-%3A-Construction-of-and-a-Teddman-Newman/c91b4dd53e25f611d2ac066f15db3ef911fb444b>, accessed on 2019-05-06.
- The R Foundation for Statistical Computing. 2016. R (Version 3.3-0). <http://www.r-project.org>, accessed on 2017-08-07.

- Thomason, Sarah G. 2001a. Contact induced typological change. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook = Sprachtypologie und sprachliche Universalien*, 1640–1648. Berlin: de Gruyter.
- Thomason, Sarah G. 2001b. *Language contact*. Edinburgh: Edinburgh University Press.
- Thomason, Sarah G. & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.
- Thrasher, Randolph H. 1974. *'Shouldn't ignore these strings': A study of conversational deletion*. Ann Arbor, MI: University of Michigan Ph.D. thesis.
- Tongue, R. K. 1979. *The English of Singapore and Malaysia*, 2nd edn. Singapore: Eastern Universities Press.
- Torres Cacoullos, Rena & Catherine E. Travis. forthc. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* (to appear 2019). https://sites.psu.edu/rct11/files/2017/08/VariationistTypology-revised_2017.8.23_DV-10g2yxo.pdf, accessed on 2019-05-06.
- Torres Cacoullos, Rena & Catherine E. Travis. 2010. Variable *yo* expression in New Mexico: English influence? In Susana Rivera-Mills & Daniel J. Villa (eds.), *Spanish of the U.S. Southwest: A language in transition*, 185–206. Madrid: Iberoamericana.
- Torres Cacoullos, Rena & 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.
- Torres Cacoullos, Rena & Catherine E. Travis. 2015. Foundations for the study of subject pronoun expression in Spanish in contact with English: Assessing interlinguistic (dis)similarity via intra-linguistic variability. In Ana M. Carvalho, Rafael Orozco & Naomi L. Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 81–100. Washington, DC: Georgetown University Press.
- Torres Cacoullos, Rena & James A. Walker. 2009. The present of the English future: Grammatical variation and collocations in discourse. *Language* 85(2). 321–354.
- Travis, Catherine E. 2005. The yo-yo effect. In Randall Gess & Edward J. Rubin (eds.), *Theoretical and experimental approaches to Romance linguistics*, 329–349. Amsterdam: John Benjamins Publishing Company.
- Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19(2). 101–135.
- Travis, Catherine E. & Amy M. Lindstrom. 2016. Different registers, different grammars? Subject expression in English conversation and narrative. *Language Variation and Change* 28(1). 103–128.
- Travis, Catherine E. & Rena Torres Cacoullos. 2012. What do subject pronouns do in discourse? Cognitive, mechanical and constructional factors in variation. *Cognitive Linguistics* 23(4). 711–748.
- Travis, Catherine E. & Rena Torres Cacoullos. 2014. Stress on *I*: Debunking unitary contrast accounts. *Studies in Language* 38(2). 372–404.
- Trudgill, Peter. 2002. *Sociolinguistic variation and change*. Edinburgh: Edinburgh University Press.
- Trudgill, Peter & Jean Hannah. 1985. *A guide to varieties of standard English*, 2nd edn. London: Hodder.
- Tsao, Feng-fu. 1977. *A functional study of topic in Chinese: The first step toward discourse analysis*. Los Angeles, California: USC Doctoral dissertation.

- Valian, Virginia. 1990. Null subjects: A problem for parameter setting models of language acquisition. *Cognition* 35(2). 105–122.
- Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40(1-2). 21–81.
- Velupillai, Viveka. 2013. Hawai'i Creole structure dataset. In Susanne M. Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole language Structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://apics-online.info/contributions/26>, accessed on 2019-05-06.
- Wagner, Susanne. 2012. *Null subjects in English: Variable rules, variable language?* Chemnitz: University of Technology Postdoctoral Thesis.
- Wagner, Susanne. 2018. Never saw one: First-person null subjects in spoken English. *English Language and Linguistics* 22(1). 1–34. Published online 18 July 2016. doi:10.1017/S1360674316000216, accessed on 2019-05-06.
- Wales, Katie. 1996. *Personal pronouns in present-day English*. New York: Cambridge University Press.
- Walkden, George. 2013. Null subjects in Old English. *Language Variation and Change* 25(2). 155–178.
- Wang, Qi, Diane Lillo-Martin, Catherine T. Best & Andrea Levitt. 1992. Null subject versus null object: Some evidence from the acquisition of Chinese and English. *Language Acquisition* 2(3). 221–254.
- Ward, Gregory & Betty J. Birner. 2004. Information structure and non-canonical syntax. In Laurence R. Horn & Gregory Ward (eds.), *The handbook of pragmatics*, 153–174. Malden, MA: Blackwell.
- Ward, Gregory, Betty J. Birner & Rodney D. Huddleston. 2002. Information packaging. In Rodney D. Huddleston, Geoffrey K. Pullum & Laurie Bauer (eds.), *The Cambridge grammar of the English language*, 1365–1447. Cambridge: Cambridge University Press.
- Wee, Lionel. 2003a. The birth of a particle: *Know* in Colloquial Singapore English. *World Englishes* 22(1). 5–13.
- Wee, Lionel. 2003b. The grammar of ignorance: The *don't know* construction. In David Deterding, Low Ee Ling & Adam Brown (eds.), *English in Singapore: Research on grammar*, 22–30. Singapore: McGraw-Hill Education.
- Wee, Lionel. 2004. Singapore English: Morphology and syntax. In Bernd Kortmann, Edgar W. Schneider, Kate Burridge & Rajend Mesthrie (eds.), *A handbook of varieties of English*, Volume 2, 1058–1072. Berlin [etc.]: Mouton de Gruyter.
- Wee, Lionel. 2014. Linguistic chutzpah and the Speak Good Singlish movement. *World Englishes* 33(1). 85–99.
- Wee, Lionel. 2018. *The Singlish controversy: Language, culture and identity in a globalizing world*. Cambridge: Cambridge University Press.
- Wee, Lionel & Umberto Ansaldi. 2004. Nouns and noun phrases. In Lisa Lim (ed.), *Singapore English: A grammatical description*, 57–74. Amsterdam: Benjamins.
- Wee, Lionel, Robbie B. H. Goh & Lisa Lim (eds.). 2013. *The politics of English: South Asia, Southeast Asia and the Asia Pacific*. Amsterdam: Benjamins.
- Weinreich, Uriel. 1963. *Languages in contact: Findings and problems*. The Hague [etc.]: Mouton Publishers.
- Weinreich, Uriel, William Labov & Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Winfred Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics*, 95–195. Austin: University of Texas Press.

- Weir, Andrew. 2012. Left-edge deletion in English and subject omission in diaries. *English Language and Linguistics* 16(1). 105–129.
- White, Lydia. 1985. The pro-drop parameter in adult Second Language Acquisition. *Language Learning* 35(1). 47–61.
- Wilson, Anna E. 2014. Testing the limits of orthogonality: A study of the interaction between lexical frequency and independent variables. *IULC Working Papers* 14(2). <https://scholarworks.iu.edu/journals/index.php/iulcwp/article/view/26209/31832>, accessed on 2019-05-06.
- Winford, Donald. 2003. *An introduction to contact linguistics*. Malden, MA [etc.]: Blackwell.
- Winford, Donald. 2017. World Englishes and creoles. In Markku Filppula, Juhani Klemola & Devyani Sharma (eds.), *The Oxford handbook of World Englishes*. Oxford [etc.]: Oxford University Press.
- Winkle, Claudia. 2015. *Non-canonical structures, they use them differently: Information packaging in spoken varieties of English*. Freiburg: Albert-Ludwigs-Universität, Universitätsbibliothek.
- Wong, Jock O. 2005. “Why you so Singlish one?” A semantic and cultural interpretation of the Singapore English particle one. *Language in Society* 34(2). 239–275.
- Wong, Jock O. 2014. *The culture of Singapore English*. Cambridge [etc.]: Cambridge University Press.
- Wong, May. 2012. Hong Kong English. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The Mouton world atlas of variation in English*, 548–561. Berlin: De Gruyter Mouton.
- Wong, May. 2013. Hong Kong English. In Bernd Kortmann & Kerstin Lunkenheimer (eds.), *The electronic world atlas of varieties of English*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://ewave-atlas.org/languages/56>, accessed on 2019-05-06.
- Wong, May. 2017. *Hong Kong English: Exploring lexicogrammar and discourse from a corpus-linguistic perspective*. London: Palgrave Macmillan.
- Wratisl, Melani & Peter Gallman (eds.). 2011. *Null pronouns*. Berlin: De Gruyter Mouton.
- Xiao, Yun. 2002. The syntactic development of school-age Chinese-speaking children learning English. *IRAL - International Review of Applied Linguistics in Language Teaching* 40(3). 235–271.
- Xu, Liejiong. 2015. Topic prominence. In Chaofen Sun & William S.-Y. Wang (eds.), *The Oxford handbook of Chinese linguistics*, 393–403. Oxford: Oxford University Press.
- Yao, Xinyue. 2016. Cleft constructions in Hong Kong English. *English World-Wide* 37(2). 197–220.
- Yip, Po-Ching & Don Rimmington. 2004. *Chinese: A comprehensive grammar*. London & New York: Routledge.
- Yip, Virginia & Stephen Matthews. 2007. *The bilingual child: Early development and language contact*. Cambridge, UK & New York: Cambridge University Press.
- Yuan, Boping. 1997. Asymmetry of null subjects and null objects in Chinese speakers’ L2 English. *Studies in Second Language Acquisition* 19(4). 467–497.
- Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Deleting named morphemes. *Lingua* 59(2-3). 155–175.

Appendix A: Data

Table A.1: GSSEC files in ICE-Singapore

GSSEC file	words	speakers				ICE file
1:1	2,149	1:1:a-I	1:1:b-C	1:1:c-C		ICE-S1A-086
1:2	807	1:2:a-I	1:2:b-C	1:2:c-?		
1:3	5,024	1:3:a-C	1:3:b-?	1:3:c-I	1:3:d-C	ICE-S1A-088
2:1	1,523	2:1:a-C	2:1:b-?	2:1:c-C		ICE-S1B-079
2:2	520	2:2:a-C	2:2:b-C			ICE-S1A-084
3:1	2,583	3:1:a-M	3:1:b-M			ICE-S1A-084
3:2	1,214	3:2:a-M	3:2:b-M	3:2:c-M		
3:3	896	3:3:a-M	3:3:b-M	3:3:c-M		
4:1	1,252	4:1:a-M	4:1:b-M	4:1:c-M		ICE-S1A-069
4:2	505	4:2:a-M	4:2:b-M			
4:3	2,069	4:3:a-M	4:3:b-M			ICE-S1A-079
4:4	447	4:4:a-M	4:4:b-M	4:4:c-C	4:4:d-C	
5:1	1,618	5:1:a-I	5:1:b-I			ICE-S1A-086
5:6	5,727	5:6:a-I	5:6:b-I	5:6:c-I		
5:8	2,350	5:8:a-I	5:8:b-I			ICE-S1A-085
6:1	8,184	6:1:a-I	6:1:b-I			
6:2	8,031	6:2:a-I	6:2:b-I	6:2:c-I	6:2:d-C	
6:3	2,142	6:3:a-I	6:3:b-I			ICE-S1A-087
7:6	452	7:6:a-C	7:6:b-C			
7:7	1,082	7:7:a-C	7:7:b-C			
8:1	3,294	8:1:a-C	8:1:b-C			ICE-S1A-090
11:2	2,242	11:2:a-I	11:2:b-I			

Speaker ethnicity: C = Chinese, M = Malay, I = Indian, ? = unknown

Appendix B: Additional statistics

Table B.1: Mixed-effects model with speaker as random intercept ICE-GB

Random effects:					
Groups name			Variance		Std.Dev.
Speaker (Intercept)			0.3440		0.5865
Number of obs: 3,468, groups: speaker, 50					
Fixed effects:	Coefficient	SE	Z	p-Value	
Intercept	-1.9356	0.2929	-6.608	<.0001	***
and: y	3.7976	0.3951	9.612	<.0001	***
Clause: subordinate	-1.2664	0.4711	-2.688	0.0072	**
Clause: question	-0.1324	0.7721	-0.172	0.8638	
Position: 2	-2.5000	0.5484	-4.559	<.0001	***
Position: 3	-0.7691	0.5957	-1.291	0.1967	.
Position: >3	-0.5793	0.3162	-1.832	0.0670	.
Person: second	-1.1210	0.4244	-2.642	0.0083	**
Person: third	-0.0574	0.2938	-0.195	0.8450	
Reference: non-referential	0.8527	0.3384	2.520	0.0117	*
Switch: partial	-0.3863	0.4489	-0.861	0.3895	
Switch: full	-0.6734	0.2820	-2.388	0.0170	*
Persistence: NP	0.6353	0.3893	1.632	0.1027	
Persistence: zero	1.4183	0.5402	2.625	0.0087	**
Verb Type: psychological	-0.6514	0.3219	-2.024	0.0430	*
Verb Type: auxiliary	-1.9198	0.2837	-6.767	<.0001	***
Verb Type: modal	-0.9615	0.3520	-2.732	0.0063	**

Table B.2: Mixed-effects model with verb token as random intercept ICE-GB

Random effects:					
Groups name			Variance		Std.Dev.
Verb Token (Intercept)			0.6965		0.8346
Number of obs: 3,468, groups: Verb Token, 148					
Fixed effects:	Coefficient	SE	Z	p-Value	
Intercept	-2.3230	0.3385	-6.862	<.0001	***
and: y	3.7248	0.3941	9.452	<.0001	***
Clause: subordinate	-1.3900	0.4860	-2.860	0.0042	**
Clause: question	-0.5475	0.7896	-0.693	0.4881	
Position: 2	-2.4929	0.5664	-4.401	<.0001	***

Table B.2: (continued)

Random effects:				
Groups name			Variance	Std.Dev.
Verb Token (Intercept)			0.6965	0.8346
Number of obs: 3,468, groups: Verb Token, 148				
Fixed effects:	Coefficient	SE	Z	p-Value
Position: 3	-0.5525	0.6090	-0.907	0.3643
Position: >3	-0.5562	0.3274	-1.699	0.0893
Person: second	-1.2277	0.4397	-2.792	0.0052
Person: third	0.2778	0.3083	0.901	0.3675
Reference: non-referential	1.0017	0.3550	2.822	0.0048
Switch: partial	-0.1469	0.4667	-0.315	0.7530
Switch: full	-0.7841	0.2888	-2.715	0.0066
Persistence: NP	0.5096	0.4054	1.257	0.2087
Persistence: zero	1.4182	0.5499	2.579	0.0099
Verb Type: psychological	-0.4673	0.4244	-1.101	0.2709
Verb Type: auxiliary	-1.2885	0.4598	-2.802	0.0051
Verb Type: modal	-0.8590	0.5079	-1.691	0.0908

Table B.3: Analysis of Deviance model comparison including factor group turn length ICE-GB

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType + Turn Length					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,451	712.14			
2	3,448	705.23	3	6.9037	0.0750.

Table B.4: Analysis of Deviance model comparison excluding factor group switch reference ICE-GB

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Persistence + VerbType					
Model 2: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,453	717.66			
2	3,451	712.14	2	5.5196	0.0633

Table B.5: Minimal adequate model ICE-GB - overview

Model: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType							
		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrimination Indexes	
Obs	3,468	LR chi2	395.45	R2	0.394	C	0.867
overt	3,338	d.f.	16	g	1.776	Dxy	0.735
zero	130	Pr(> chi2)	<0.0001	gr	5.906	gamma	0.747
max deriv 8e-10				gp	0.056	tau-a	0.053
				Brier	0.024		

Table B.6: Model evaluation by bootstrapping minimal adequate model ICE-GB (n = 142)

	index.orig	training	test	optimism	index corrected
Dxy	0.7385	0.7567	0.7274	0.0292	0.7093
R2	0.3938	0.4025	0.3794	0.0231	0.3707
Intercept	0.0000	0.0000	-0.1062	0.1062	-0.1062
Slope	1.0000	1.0000	0.9559	0.0441	0.9559
Emax	0.0000	0.0000	0.0311	0.0311	0.0311
D	0.1137	0.1167	0.1093	0.0073	0.1064
U	-0.0006	-0.0006	0.0003	-0.0009	0.0003
Q	0.1143	0.1172	0.1090	0.0082	0.1061
B	0.0242	0.0242	0.0247	-0.0005	0.0247
g	1.7760	1.9195	1.8263	0.0932	1.6828
gp	0.0558	0.0568	0.0555	0.0013	0.0546

Table B.7: Analysis of Deviance model comparison with monofactorial model “and” ICE-GB

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,451	712.14			
2	3,466	844.40	-15	-132.26	<2.2e-16 ***

Table B.8: Logistic regression model including factor group VP complexity ICE-GB

	Coefficient	SE	Z	p-Value	
Intercept	-2.7117	0.2743	-9.885	<2e-16	***
and: y	3.7605	0.3541	10.619	<2e-16	***
Clause: subordinate	-1.1174	0.4551	-2.455	0.0141	*
Clause: question	-0.6450	0.7600	-0.855	0.3924	
Position: 2	-2.6100	0.5462	-4.778	1.77e-06	***
Position: 3	-0.7011	0.5700	-1.230	0.2187	
Position: >3	.04706	0.3100	-1.564	0.1179	
Person: second	-0.9332	0.3914	-2.384	0.0171	*
Person: third	-0.2015	0.2774	-0.726	0.4676	
Reference: non-referential	0.5300	0.3231	1.640	0.1009	
Switch: partial	-0.4800	0.4296	-1.117	0.2638	
Switch: full	-0.7577	0.2700	-2.807	0.0050	**
Persistence: NP	0.4934	0.3766	1.310	0.1901	
Persistence: zero	1.4192	0.4958	2.863	0.0042	**
VPcomplexity: 2	-0.1635	0.2378	-0.688	0.4916	
VPcomplexity: 3+	-0.0452	0.3423	-0.132	0.8951	

Table B.9: Wald statistics of factor group significance including factor group VP complexity ICE-GB

	Df	Deviance	Resid.Df	Resid.Dev.	Pr(>Chi)	
NULL	3467	1108.85				
and	1	264.451	3,466	844.40	< 2.2e-16	***
Clause	2	18.222	3,464	826.18	0.0001	***
Position	3	41.977	3,461	784.20	4.058e-09	***
Person	2	5.553	3,459	778.65	0.0623	.
Reference	1	0.191	3,458	778.46	0.6624	
Switch	2	8.559	3,456	769.90	0.0138	*
Persistence	2	8.532	3,454	761.37	0.0140	*
VPcomplexity	2	0.493	3,452	760.87	0.7814	

Table B.10: Analysis of Deviance model comparison including factor group VP complexity ICE-GB

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VP complexity					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,451	712.14			
2	3,452	760.87	-1	-48.733	2.933e-12 ***

Table B.11: Logistic regression model including interaction term Person:VerbType ICE-GB

	Coefficient	SE	Z	p-Value	
Intercept	-1.2354	0.4951	-2.495	0.0126	*
and: y	3.9008	0.3857	10.113	<2e-16	***
Clause: subordinate	0.1411	0.7624	0.185	0.8531	
Clause: question	-1.5228	0.4736	-3.215	0.0013	**
Position: 2	-2.6480	0.5486	-4.827	1.39e-06	***
Position: 3	-0.7647	0.6010	-1.272	0.2033	
Position: >3	-0.5146	0.3098	-1.661	0.0968	.
Person: second	-1.9276	0.6954	-2.772	0.0056	**
Person: third	0.5472	0.4096	1.336	0.1816	
Reference: non-referential	-1.1410	0.3436	-3.321	0.0009	***
Switch: partial	-0.3064	0.4475	-0.685	0.4936	
Switch: full	-0.6218	0.2814	-2.209	0.0271	*
Persistence: NP	0.7754	0.3918	1.979	0.0478	*
Persistence: zero	1.3996	0.5314	2.634	0.0084	**
VerbType: psychological	-0.8329	0.4994	-1.668	0.0954	.
VerbType: auxiliary	-0.7507	0.4622	-1.624	0.1044	
VerbType: modal	-0.4089	0.5420	-0.754	0.4506	
Person2:VerbTypeP	3.3038	0.9190	3.595	0.0003	***
Person3:VerbTypeP	-0.6074	0.7729	-0.786	0.4319	
Person2:VerbTypeX	-14.476	529.99	-0.027	0.9782	
Person3:VerbTypeX	-1.8526	0.5957	-3.110	0.0019	**
Person2:VerbTypeM	0.9021	1.0763	0.838	0.4020	
Person3:VerbTypeM	-1.3686	0.7803	-1.754	0.0794	.

Table B.12: Analysis of Deviance model comparison including interaction term Person:VerbType ICE-GB

Analysis of Deviance Table						
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType						
Model 2: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType + Person * VerbType						
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)	
1	3,451	712.14				
2	3,445	680.09	6	32.056	1.592e-05	***

Table B.13: Minimal adequate model first person ICE-GB - overview

Model: S ~ and + Clause + Persistence + VerbType							
		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrimination Indexes	
Obs	901	LR chi2	95.11	R2	0.310	C	0.833
overt	857	d.f.	7	g	1.061	Dxy	0.666
zero	44	Pr(> chi2)	<0.0001	gr	2.889	gamma	0.730
max deriv 1e-5				gp	0.058	tau-a	0.062
				Brier	0.035		

Table B.14: Model evaluation by bootstrapping minimal adequate model first person ICE-GB (n = 169)

	index.orig	training	test	optimism	index corrected
Dxy	0.6673	0.6574	0.6181	0.0394	0.6280
R2	0.3101	0.3241	0.2746	0.0495	0.2606
Intercept	0.0000	0.0000	-0.2185	0.2185	-0.2185
Slope	1.0000	1.0000	0.9067	0.0933	0.9067
Emax	0.0000	0.0000	0.0659	0.0659	0.0659
D	0.1045	0.1103	0.0918	0.0185	0.0860
U	-0.0022	-0.0022	0.0015	-0.0037	0.0015
Q	0.1067	0.1125	0.0904	0.0221	0.0845
B	0.0352	0.0350	0.0361	-0.0011	0.0363
g	1.0609	1.6576	1.4481	0.2095	0.8514
gp	0.0583	0.0608	0.0557	0.0051	0.0533

Table B.15: High-frequency verb lemmas

GB		IN		HK		SG	
be	1,295	be	1,184	be	1,130	be	963
do	374	have	446	have	476	do	426
have	357	do	229	do	470	will	261
would	136	can	145	think	275	can	229
can	149	will	134	can	175	have	195
think	120	should	88	will	131	get	147
will	85	think	86	want	78	go	147
go	68	go	73	go	74	think	109
say	62	get	72	would	61	want	108
get	50	come	66	get	56	say	102
know	48	say	51	know	56	know	57
could	37	know	50	like	48	come	53
want	31	mean	47	need	45	should	46
mean	27	want	45	say	44		
see	20	would	43	should	33		
		could	41				

Table B.16: Minimal adequate model ICE-IN

	Coefficient	SE	Z	p-Value	
Intercept	-2.3780	0.1818	-13.080	<2e-16	***
and: y	2.6044	0.3554	7.328	2.34e-13	***
Clause: question	-0.6330	0.6467	-0.979	0.3276	
Clause: subordinate	-2.5096	0.4193	-5.985	2.17e-09	***
Person: second	0.2176	0.3296	0.660	0.5091	
Person: third	1.0768	0.2165	4.975	6.54e-07	***
Switch: partial	-2.2550	0.6242	-3.613	0.0003	***
Switch: full	-0.0159	0.2095	-0.076	0.9395	
Persistence: NP	1.3627	0.3002	4.539	5.65e-06	***
Persistence: zero	2.9699	0.3410	8.708	<2e-16	***
Verb Type: psychological	-1.1126	0.2900	-3.837	0.0001	***
Verb Type: auxiliary	-2.8724	0.2791	-10.293	<2e-16	***
Verb Type: modal	-2.9426	0.4502	-6.536	6.32e-11	***

Table B.17: Analysis of Deviance model comparison minimal adequate model ICE-IN

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and + Clause + Person + Switch + Persistence + VerbType					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,619	925.20			
2	3,623	931.25	-4	-6.0525	0.1953

Table B.18: Model evaluation by bootstrapping logistic regression model ICE-IN (n = 125)

	index.orig	training	test	optimism	index corrected
Dxy	0.8051	0.8174	0.7999	0.0175	0.7876
R2	0.4148	0.4233	0.4028	0.0205	0.3942
Intercept	0.0000	0.0000	-0.0536	0.0536	-0.0536
Slope	1.0000	1.0000	0.9528	0.0472	0.9528
Emax	0.0000	0.0000	0.0203	0.0203	0.0203
D	0.1475	0.1483	0.1429	0.0055	0.1420
U	-0.0006	-0.0006	0.0003	-0.0008	0.0003
Q	0.1480	0.1489	0.1426	0.0063	0.1417
B	0.0339	0.0329	0.0343	-0.0014	0.0353
g	2.2199	2.3919	2.2680	0.1239	2.0959
gp	0.0783	0.0773	0.0776	-0.0003	0.0787

Table B.19: Minimal adequate model ICE-HK

	Coefficient	SE	Z	p-Value	
Intercept	-2.8071	0.1859	-15.100	<2e-16	***
and: y	2.5769	0.2730	9.441	<2e-16	***
Clause: question	-1.5245	0.5410	-2.818	0.0048	**
Clause: subordinate	-0.3365	0.2083	-1.615	0.1063	
Person: second	0.7776	0.2622	2.966	0.0030	**
Person: third	1.5084	0.1958	7.704	1.32e-14	***
Switch: partial	-0.8514	0.3456	-2.464	0.0138	*
Switch: full	-0.9339	0.1969	-4.742	2.12e-06	***
Persistence: NP	0.5440	0.2237	2.432	0.0150	*
Persistence: zero	1.5636	0.3114	5.021	5.13e-07	***
Verb Type: psychological	-0.5808	0.2271	-2.557	0.0106	*
Verb Type: auxiliary	-1.4639	0.2035	-7.195	6.24e-13	***
Verb Type: modal	-1.2989	0.2775	-4.681	2.86e-06	***

Table B.20: Analysis of Deviance model comparison minimal adequate model ICE-HK

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and + Clause + Person + Switch + Persistence + VerbType					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,962	1,233.8			
2	3,966	1,235.7	-4	-11.941	0.0178 *

Table B.21: Model evaluation by bootstrapping logistic regression model ICE-HK (n = 191)

	index.orig	training	test	optimism	index corrected
Dxy	0.6732	0.6822	0.6594	0.0228	0.6505
R2	0.2840	0.2929	0.2731	0.0198	0.2642
Intercept	0.0000	0.0000	-0.0985	0.0985	-0.0985
Slope	1.0000	1.0000	0.9552	0.0448	0.9552
Emax	0.0000	0.0000	0.0294	0.0294	0.0294
D	0.0996	0.1030	0.0956	0.0074	0.0922
U	-0.0005	-0.0005	0.0002	-0.0007	0.0002
Q	0.1001	0.1035	0.0954	0.0081	0.0920
B	0.0395	0.0391	0.0399	-0.0008	0.0403
g	1.3819	1.4736	1.3995	0.0741	1.3077
gp	0.0656	0.0668	0.0647	0.0020	0.0636

Table B.22: Mixed-effects model with ethnicity as random intercept GSSEC

Random effects:					
Groups name				Variance	Std.Dev.
ethnicity (Intercept)				4e-14	2e-07
Number of obs: 3,939, groups: ethnicity, 3					
Fixed effects:	Coefficient	SE	Z	p-Value	
Intercept	-1.6706	0.1476	-11.316	<2e-16	***
and: y	2.7616	0.2892	9.549	<2e-16	***
Clause: subordinate	-0.6886	0.1971	-3.493	0.0005	***
Clause: question	0.6531	0.1911	3.418	0.0006	***
Position: 2	-0.6359	0.1466	-4.337	1.44e-05	***
Position: 3	-0.9595	0.2170	-4.421	9.80e-06	***
Position: >3	-0.4208	0.1389	-3.029	0.0025	**
Person: second	-0.3332	0.1588	2.099	0.0358	*

Table B.22: (continued)

Random effects:					
Groups name				Variance	Std.Dev.
ethnicity (Intercept)				4e-14	2e-07
Number of obs: 3,939, groups: ethnicity, 3					
Fixed effects:	Coefficient	SE	Z	p-Value	
Person: third	-0.2539	0.1389	-1.828	0.0676	.
Reference: non-referential	0.8381	0.1477	5.674	1.39e-08	***
Switch: partial	-0.0534	0.2413	-0.221	0.8248	
Switch: maintenance	0.2934	0.1143	2.567	0.0102	*
Persistence: NP	0.3976	0.1759	2.260	0.0238	*
Persistence: zero	0.9681	0.1925	5.028	4.96e-07	***
Verb Type: psychological	-0.5643	0.1665	-3.389	0.0007	***
Verb Type: auxiliary	-0.7828	0.1352	-5.789	7.08e-09	***
Verb Type: modal	0.0047	0.1426	0.033	0.9734	

Table B.23: Minimal adequate model GSSEC

	Coefficient	SE	Z	p-Value	
Intercept	-1.5231	0.1238	-12.305	<2e-16	***
and: y	2.7188	0.2868	9.481	<2e-16	***
Clause: question	0.5359	0.1799	2.978	0.0029	**
Clause: subordinate	-0.7117	0.1966	-3.621	0.0003	***
Position: 2	-0.6529	0.1463	-4.462	8.10e-06	***
Position: 3	-0.9652	0.2169	-4.451	8.56e-06	***
Position: >3	-0.4318	0.1387	-3.114	0.0018	**
Reference: non-referential	0.6971	0.1286	5.422	5.89e-08	***
Switch: partial	-0.3361	0.2379	-1.413	0.1577	
Switch: full	-0.2689	0.1137	-2.365	0.0180	*
Persistence: NP	0.3475	0.1719	2.021	0.0433	*
Persistence: zero	0.9787	0.1922	5.093	3.52e-07	***
Verb Type: psychological	-0.5469	0.1661	-3.293	0.0010	***
Verb Type: auxiliary	-0.7562	0.1337	-5.654	1.56e-08	***
Verb Type: modal	0.0130	0.1421	0.092	0.9270	

Table B.24: Analysis of Deviance model comparison minimal adequate model GSSEC

Analysis of Deviance Table					
Model 1: S ~ and + Clause + Position + Person + Reference + Switch + Persistence + VerbType					
Model 2: S ~ and + Clause + Position + Reference + Switch + Persistence + VerbType					
	Resid. Df	Resid. Dev.	Df	Deviance	Pr(>Chi)
1	3,922	2,586.4			
2	3,924	2,591.5	-2	-5.1225	0.0772

Table B.25: Model evaluation by bootstrapping logistic regression model GSSEC (n = 200)

	index.orig	training	test	optimism	index corrected
Dxy	0.4068	0.4190	0.3950	0.0241	0.3828
R2	0.1351	0.1421	0.1284	0.0136	0.1214
Intercept	0.0000	0.0000	-0.1046	0.1046	-0.1046
Slope	1.0000	1.0000	0.9449	0.0551	0.9449
Emax	0.0000	0.0000	0.0327	0.0327	0.0327
D	0.0722	0.0763	0.0685	0.0077	0.0645
U	-0.0005	-0.0005	0.0003	-0.0008	0.0003
Q	0.0727	0.0768	0.0683	0.0085	0.0642
B	0.0943	0.0941	0.0948	-0.0007	0.0950
g	0.8624	0.9044	0.8515	0.0528	0.8096
gp	0.0879	0.0908	0.0863	0.0045	0.0833

Index

- African American Vernacular English 145
age 6, 33, 60, 90, 173, 186, 197, 222
analytic language 2, 24, 100, 105, 111, 143
Angloversal 121, 141
animate 51–52, 110
APiCS 117, 123
areality 100, 121, 141, 233, 248
auxiliary verb 56, 81, 150, 182, 185, 193, 205, 216, 227
- bare conditional 140, 152, 220, 233
BE deletion 137, 143, 150, 248
Bislama 26, 53, 118, 123, 146, 156, 179
bootstrapping 69, 86, 199
borrowing scale 113–114, 120
Butler English 142
- Cantonese 7, 26, 33, 130–132, 136, 139
Chinese Pidgin English 118, 130
cleft construction 20, 128, 133, 144, 156
collocation 43, 61–63, 74, 77, 90, 144, 147, 164–170, 190, 223, 244, 246, 250
collostructional nativisation 144, 223, 248
comment 19, 104–107, 128, 210, 213
complexity 32, 43, 56, 245
– verb phrase complexity 41, 55, 81, 94, 237, 246
concordance index 69, 86, 91, 176, 185, 188, 196–197, 199, 209
conditional clause 49, 140, 152
convergence 4, 32, 38, 113, 126, 128, 144, 147, 155, 231, 234, 249, 251
creole 3, 25, 111, 115–116, 142, 145, 156, 179, 191
creoloid 115
- diary 2, 36, 49
diglossia 135
dummy pronoun 18, 37, 52, 106, 123, 127, 129, 133, 143, 202, 233
Dynamic model 4, 10, 120–121, 125, 128, 130, 134, 144, 219, 222, 233, 248, 250
- education 5–6, 28, 124, 130, 134, 137
- EFL 7, 120
ENL 120
ESL 7, 120
ethnicity 7, 121, 130, 134, 137, 144, 198, 219, 222, 246
eWAVE 120, 122, 129, 133, 135, 137, 233
existential *got* 137, 139, 165, 167, 204, 212, 217, 220, 249–250
exonormative 121, 132
expletive 18, 22, 24–25, 225, 249
- feature pool 66, 115, 120, 144, 248, 250
Finnish 2, 25, 31–32, 91
focus 2, 19–22, 109, 128, 202
founder effect 136
- gender 4, 6, 28, 33, 42, 59, 110, 173, 186, 197–198, 222
generic reference 17, 26, 51–52, 76–77, 92, 101, 109, 144, 156, 213, 218, 225, 231, 240
German 25
GloWbE 5
grammaticalisation 112, 124, 220, 249
- Hindi 125–126, 144, 175, 234
Hokkien 131, 136, 139
- idiolect 218
imperative 34, 45, 104, 110, 230, 244, 249
inanimate 17, 110, 115–116
Indo-European 104, 126
information packaging *see* information structure
information structure 9, 16, 18–22, 247
intonation 9, 22, 49, 137, 220, 246
Italian 1, 23, 25–26, 33, 100
- Kannada 128, 175
- language acquisition 3, 36, 41, 115, 132, 250
language policy 4, 125, 131, 134
Lexifier Filter Model 113, 249

- Malay 134–137, 198, 222
 – Baba Malay 136, 249
 – Bazaar Malay 136, 249
 Malayalam 126, 144, 175, 234
 Malaysian English 123, 142–143, 249
 modal verb 41, 56, 63, 77, 82, 89, 92, 140, 145, 159, 165, 168, 181, 185, 190, 193, 196, 206, 214, 226, 237, 239, 249
 model diagnostics 86, 91, 209, 229, 236, 243
 Model Likelihood Ratio Test 69, 86, 176, 185, 188, 196, 199, 209
 mother tongue 6–7, 125, 134
 multicollinearity 70, 176, 188, 199
 multilingualism 113, 120, 125–126, 135–136, 174, 219, 251
- narrative 6, 9, 16, 22, 31, 41, 111, 126, 167, 183, 211, 247
 nativisation 10, 120, 125, 128, 131, 134, 144, 187, 223, 227, 233, 246–250
 Newfoundland English 2, 38, 41, 47, 56, 58, 90, 92, 94, 237
 null object 25, 103, 117, 129, 206, 217, 233
 null relative pronoun 132, 138, 142, 152, 190, 228, 251
- orderly heterogeneity 4, 27
- Pakistani English 142
 parameter 9, 23, 36, 104, 111, 120, 124, 133, 225, 243, 250
 Principles and Parameters 1, 25
 Pear Stories 22, 40, 110, 247
 Peranakan 136
 persistence 31, 42, 54, 64, 66, 79–80, 86, 89, 92, 109, 147, 158, 180, 185, 191, 196, 205, 209, 212, 226, 229, 237, 240, 242, 245
 priming 31–32, 39, 42, 54, 80, 93, 180, 185, 192, 218, 238, 245
 Philippine English 142
 phonology 21, 28, 36, 114, 121, 128, 131, 137, 157, 220, 246
 pidgin 115, 118, 120, 136, 142
 Portuguese 25, 31, 100
 pragmatics 18, 33, 99, 103, 106, 115, 117, 128–129, 213, 215, 217
- predicative adjective 113, 150, 216, 232, 244, 249
 processing 19, 33, 41, 53, 56, 64, 237, 245
 pro-drop 2, 15, 23–26, 36, 100–104, 117–119, 142–143, 225, 250
 proficiency 7, 116, 120, 131, 135, 137, 142, 173, 187, 218, 220, 222, 250
 psychological verb 30, 32, 56, 77, 81, 89, 147, 159, 168, 185, 196, 209, 237, 240
- question 46, 49, 64, 75, 84, 89, 110, 121, 137, 141, 144, 153, 155, 178, 185, 190, 196, 201–202, 208, 218, 220, 227–228, 238, 250
 quotative verb 40, 48, 73, 152, 167, 171, 178
- R^2 69, 86, 176, 185, 188, 196, 199, 209
 radical pro-drop 24, 100–103
 random effects 32, 42, 67, 82, 90, 198, 221, 244
 reduplication 128
 register 2, 8, 16, 36, 49, 57
 relative clause 22, 36, 49, 131, 137, 142, 144, 152, 190, 197, 214, 221, 233
 relative particle 138, 214, 220
 repeated name penalty 245
 replication 112, 137, 146, 152, 214, 229, 249
 rheme 19
- second language acquisition 111, 115, 132
 simplification 115, 132, 143
 Sinitic 100, 135, 150, 153, 249
 South African Indian English 123, 173, 238
 Spanish 1, 6, 9, 15, 23–25, 29–32, 36, 38–40, 47, 53, 56, 89, 100, 103, 110, 116, 232, 238, 245, 247, 252
 Speak Good English movement 135, 251
 Sprachbund 126, 129
 Sri Lankan English 122, 142
 Standard Average European 2, 24
 style 9, 47, 137, 141, 149, 151, 222
 – acrolect 123, 173, 225, 238
 – basilect 123, 225, 238, 249
 – mesolect 123, 238
 – stylistic variation 137, 159, 220, 246
 switch reference 29, 33, 40–41, 42, 53–54, 64, 68, 78–79, 84, 86, 91–92, 109–110,

- 158, 179–180, 185, 191, 196, 205, 209, 212, 220, 226, 231, 236, 240, 245
- Tamil 134, 175
- Teochew 131, 136, 138
- theme 19
- Three Circles model 120
- topic
 - topicality scale 20, 22, 246
 - topic-drop 26, 36
 - topic prominence 100, 104–109, 113, 117, 127, 133, 140, 148, 209–217, 220, 222, 249
- topicalisation 105, 127, 210, 218
- VarbRul 66, 71, 178
- variable context 10, 27, 29, 33, 39, 41, 44, 57, 92, 94, 110, 145, 150, 185, 195, 208, 221, 229, 234, 240, 244, 251
- variance inflation factor 70, 87, 176, 188, 199
- varioversal 121, 141
- Wald statistics 67, 85, 176, 187, 199, 223, 235
- WALS 100, 118, 123
- X or not* question 138, 141, 153, 202, 220, 228, 250
- yes-no* question 138, 153
- Zipf's law 56

