



**Cross-generational linguistic variation in the  
Canberra Vietnamese heritage language community:  
A corpus-centred investigation**

NGUYEN HOANG BAO NGUYEN



Churchill College

This dissertation is submitted for the degree of Doctor of Philosophy

October, 2020



## DECLARATION

---

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit of 80,000 words.

NGUYEN HOANG BAO NGUYEN

October, 2020



# ABSTRACT

---

## **Cross-generational linguistic variation in the Canberra Vietnamese heritage language community: a corpus-centred investigation**

**Nguyen H.B. Nguyen**

This dissertation investigates cross-generational linguistic differences in the Canberra Vietnamese bilingual community, with a particular focus on Vietnamese as the heritage language. Specifically, it documents the vernacular and considers key aspects of this data from different theoretical perspectives. Its main contribution is an insight into a rarely-studied heritage language variety in a contact community that has never been examined.

The dissertation consists of five core chapters, organised into two parts. In the first part (Chapters 2–3), I describe how I documented the vernacular and created the Canberra Vietnamese English Corpus (CanVEC), an original corpus compiled specifically for this study that is also the first to be freely available for research purposes. The corpus consists of over ten hours of spontaneous speech produced by 45 Vietnamese-English bilingual speakers across two generations living in Canberra. In the second part of the study (Chapters 4–6), I put the corpus to use and investigate aspects of the cross-generational differences in Vietnamese as the heritage language in this community.

In particular, I first probe the Vietnamese heritage language via its participation in the code-switching discourse (Chapter 4). In doing so, I focus on the applicability of the Matrix Language Framework (MLF) (Myers-Scotton, 1993, 2002) and its associated Matrix Language (ML) Turnover Hypothesis (Myers-Scotton, 1998) to the code-switching data in CanVEC. Since support for this prominent model has mainly come from language pairs that have different clausal word order or vastly different inventories of inflectional morphology, Vietnamese-English as a pair in which both languages are SVO and essentially isolating offers a tantalising testing ground for its application. Results show that the universal claims of this model do not hold so straightforwardly. CanVEC data challenges several assumptions of the MLF, with the model ultimately only being able to account for around half of the CanVEC code-switching data. I further demonstrate that even when the ML is putatively identifiable and a cross-generational ML ‘turnover’ is quantitatively observed, the predictions do not reflect the direction of structural influence that we see in CanVEC. The MLF approach therefore sheds only limited light on cross-generational language shift and variation in this community.

Given that null elements emerge as a distinct area of difficulty in Chapter 4, I take this aspect as the focal point for the next part of the investigation (Chapter 5), where I use the variationist approach (Labov, 1972 et seq.) to explore three cases where null and overt realisation alternates in Vietnamese: subjects, objects, and copulas. In doing so, I move away from the bilingual portion of CanVEC to examine the monolingual heritage Vietnamese subset directly. Results show that Vietnamese null subjects vary significantly across generations, while null objects and copulas remain stable in terms of use. As speakers also overwhelmingly prefer overt forms over null forms (~70:30) across all the three of the variables of interest, I appeal to the generative interface-oriented approach (Sorace & Filiaci, 2006 et seq.) to next examine the distribution of overt subjects, objects, copulas (Chapter 6). These results converge with what was found for null forms: cross-generational effects were observed for pronominal subjects, but not pronominal objects and copulas. This finding also supports the importance of a distinction drawn in previous works between internal (syntax-semantics) and external (syntax-discourse/pragmatics) interface phenomena, with the latter being seemingly more susceptible to change.

Ultimately, this dissertation highlights the empirical and theoretical value of studying rarely-considered contact varieties, while deploying an integrated approach that acknowledges the multifaceted complexity of the contact communities where these varieties are spoken.

## ACKNOWLEDGEMENTS

---

I would first like to thank my supervisors, Theresa Biberauer and Henriëtte Hendriks, for seeing this dissertation through to completion, and Margaret Deuchar for enabling some parts of the project. Henriëtte admitted me into the programme and gave me an opportunity to pursue this doctorate. I am also grateful for her feedback at the culminating stage, which led to some key improvements in the final presentation. Theresa in particular stepped in at a very difficult point in the process, and without her, this PhD would have never seen the light of day. While her broad expertise and academic rigour may be well-known to many in the field, I count myself lucky to have also experienced first-hand her generosity towards her students. Not only did she spend countless hours reading and criticising various drafts which form the heart of this dissertation, pointing out problems and suggesting solutions, she also went beyond her call of duty to instil confidence and provide selfless support when various semi-academic issues arose. I am deeply indebted to her for advancing my thinking and execution of this work, while still giving me ample space to freely explore and produce research that faithfully reflects my own intellect and conclusions. My growth as a researcher is a result of Theresa's extraordinary hard work.

I would also like to thank Jane Simpson and Catherine Travis at the Australian National University, where I had the best time before Cambridge, for setting me on a course of doing bigger and better things with Linguistics. Had Jane not accepted me into the Linguistics programme at ANU despite my unrelated BA, introduced me to all areas of Linguistics, and then wrote my references for Cambridge, I would not be where I am today. Catherine co-supervised my Masters with Jane, and I am grateful that she continued to put up with my questions even after her duty of supervision had long expired. [Chapter 5](#) in particular was in part inspired by her work on null subjects in the New Mexico community.

I next want to thank all my informants and the speakers of CanVEC without whom this dissertation would not have been written. Many thanks also go to Han Sloetjes at the Max Planck Institute for Psycholinguistics for his technical support with alignment in ELAN; to Thomas Marge from the Cambridge Centre for Mathematical Sciences for his personal tutorials on statistics; and to Marc Brunelle from the University of Ottawa and Lương Xuân Vũ and Linh Tuyển Hoàng from the Vietnamese Lexicography Centre (Vietlex) for providing me with monolingual colloquial Vietnamese materials to use as examples in multiple places in this dissertation. I would also like to thank many other linguist friends and colleagues who taught me valuable things and provided feedback at different stages: Josh Brown, Andrew Caines, Alexander Cairncross, Draško Kaščélan, Sana Kidwai, Vicky Lee, Oliver Mayeux, Tom Meadows, Eleonora Serra, Julio Song, and Kayeon Yoo. I thank Alex and Julio in particular for being especially enthusiastic and crit-

ical readers of the entire manuscript. A special thank-you also goes out to all my friends of the Last-minute Proofreading Taskforce, who patiently crossed out all my extra words and put all my agreements in order. Mariano Felice in particular deserves a special mention for handling my requests for *LaTeX* typesetting support at all hours, and for always spotting the tiniest missing comma in my code. Finally, my deepest gratitude goes to my examiners, Paula Buttery and Pieter Muysken, for an enjoyable viva and for their encouragingly constructive feedback.

I was able to conduct this PhD research thanks to full funding from the Cambridge International Trust. Together with Churchill College and the University Hardship Fund, the Trust was particularly generous in extending my scholarship into my fourth year. The Philological Society and the MMLL Faculty also funded my fieldwork and several conference travels. Rebecca Sawalmeh and Barry Phipps at Churchill College also went out of their way to give me pastoral care and practical support with many funding applications. Importantly, I would also like to acknowledge the various grants made accessible to me by Theresa, which allowed me to fully concentrate on the write-up at the final stage.

Time in Cambridge is not all about the PhD, and in fact, time in Cambridge would not have been sustainable *at all* without my good friends, their love and humour, especially at difficult times. Among some of the first friends that I made in the U.K, I would like to thank Simon Corkery, Rosa Hodgkin, Jack Hodgkinson, Lachlan Lancaster, Mike Meaney, Tahmina Seddiqi, and the other Littlies for making my time at 64 the best transition. I am also indebted to Sana's Steady Support, whose unoriginal 'inspirational quotes' have somehow kept me amused during the darkest hours of this work. Among all, Kayeon Yoo deserves the most special mention, not only for sharing the funnest sleepovers and the yummiest Korean food, but also for being and remaining my most steadfast and reliable friend since day one of the PhD. Finally, I should also thank my *bạn tù kiếp trước* Oliver Mayeux for indulging my love of Karaoke during times of stress, and for the many meaningful conversations over the years. All of you have made Cambridge magical. Halfway across the globe, I cannot forget Isabelle Morgan, Henna Chhabra, Yến Nguyễn, An Vũ, Taryn Johnson and Anna Whitton for all the messages, cards, gifts and care packages that I often received. Although it has been tough not being able to hang out with you more, I hope the sacrifices are all worth it.

I also feel especially fortunate to have my partner, Christopher Bryant, along with me on this journey. I cannot begin to thank him enough, not only for the collaboration which shapes some parts of this dissertation (Nguyen & Bryant, 2020), but also for enriching my life outside Linguistics. Who would have thought that our serendipitous collaboration a few years ago would not end there, but would go on to form a rewarding *dvandva*. I look forward to our next chapter together, and to processing more corpora with you.

Finally and importantly, I would like to thank my family in Vietnam and Australia. I'm forever indebted to my parents for beating the odds to ensure that I had the best education I could have. I would also like to send a special thank-you to my sister Thảo, her husband Thi, and my beautiful nephew and nieces Minh, Mai, and Mi, for always offering me a home in Canberra. This dissertation is as much their creation as it is mine.

# CONTENTS

---

<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xviii</b>
<b>Abbreviations</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Setting the stage . . . . .	1
1.2 Research components . . . . .	3
1.2.1 Data: Introducing the Canberra Vietnamese-English Corpus . . . . .	3
1.2.2 Theoretical frameworks . . . . .	4
1.3 Overview of the study . . . . .	6
<b>PART I Documenting the Canberra Vietnamese community vernacular</b>	
<b>2 Characterising the community: Vietnamese in Canberra</b>	<b>11</b>
2.1 Introduction . . . . .	11
2.2 Vietnamese in Australia: Political history and language use . . . . .	11
2.3 The Vietnamese community in Canberra . . . . .	16
2.3.1 Defining a ‘speech community’ for Vietnamese speakers in Canberra . . . . .	17
2.3.2 Canberra Vietnamese as a community of practice . . . . .	19
2.3.3 Summary . . . . .	21
2.4 CanVEC speakers: Who are they? . . . . .	22
2.4.1 Pooling the sample . . . . .	22
2.4.2 Demographic profile and generation membership . . . . .	23
2.4.3 Social network . . . . .	27
2.4.4 Language maintenance, language attitude and language preference . . . . .	29
2.5 Chapter summary . . . . .	32

<b>3</b>	<b>Building the Canberra Vietnamese-English Corpus (CanVEC)</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Building CanVEC . . . . .	33
3.2.1	Recording procedures . . . . .	33
3.2.2	Questionnaire . . . . .	35
3.3	Annotating CanVEC . . . . .	36
3.3.1	Transcription method . . . . .	37
3.3.1.1	Sound to text . . . . .	37
3.3.1.2	Segmentation: Unit of analysis . . . . .	40
3.3.1.3	Ethical considerations . . . . .	45
3.3.2	Semi-automatic data processing . . . . .	47
3.3.2.1	Automatic language marking and Part of Speech (POS) tagging	48
3.3.2.2	Manual verification: Language-neutral items, non-linguistic items, and established borrowing . . . . .	51
3.3.2.3	Automatic translation . . . . .	53
3.3.3	Evaluation . . . . .	54
3.3.3.1	Language marking and POS-tagging . . . . .	54
3.3.3.2	Translation . . . . .	57
3.3.4	Summary . . . . .	59
3.4	Chapter summary . . . . .	59

## **PART II Cross-generational variation in the Vietnamese heritage language of the Canberra Vietnamese community**

<b>4</b>	<b>The Matrix Language in the community</b>	<b>63</b>
4.1	Introduction . . . . .	63
4.2	The Matrix Language and Matrix Language Turnover . . . . .	64
4.2.1	Myers-Scotton's Matrix Language Framework (MLF) . . . . .	64
4.2.1.1	The Content-System Morpheme distinction . . . . .	65
4.2.1.2	The Matrix Language-Embedded Language distinction . . . . .	67
4.2.2	The Matrix Language Turnover Hypothesis . . . . .	69
4.3	Application of the MLF in the literature . . . . .	72
4.3.1	Previous work using the MLF . . . . .	72
4.3.1.1	Predictive power of the MLF . . . . .	72
4.3.1.2	The MLF in 'inconvenient' language pairs . . . . .	76

4.3.2	Previous work on the Matrix Language Turnover Hypothesis . . . . .	79
4.4	Establishing the Matrix Language in CanVEC . . . . .	84
4.4.1	The System Morpheme Principle . . . . .	84
4.4.2	The Morpheme Order Principle . . . . .	87
4.4.2.1	Morpheme order within the nominal domain . . . . .	88
4.4.2.2	Polar questions . . . . .	89
4.4.2.3	Wh-questions . . . . .	89
4.4.3	Results . . . . .	90
4.5	Difficult data . . . . .	92
4.5.1	Whose Matrix Language is the Matrix Language? . . . . .	92
4.5.2	Composite Matrix Language . . . . .	94
4.5.3	Clauses with null elements . . . . .	96
4.5.4	A note on Wang's additional principles . . . . .	101
4.5.5	Summary . . . . .	105
4.6	Matrix Language Turnover in the community . . . . .	106
4.6.1	Is there a Matrix Language Turnover? . . . . .	106
4.6.2	Direction of structural borrowing . . . . .	108
4.6.3	Early syntactic knowledge: A case of stability . . . . .	111
4.6.4	A note on 'stable bilingualism' . . . . .	115
4.7	Chapter summary . . . . .	116
<b>5</b>	<b>Characterising generational differences: A variationist study</b>	<b>119</b>
5.1	Introduction . . . . .	119
5.2	Key principles of the variationist approach . . . . .	120
5.2.1	Orderly heterogeneity . . . . .	120
5.2.1.1	Orderly heterogeneity in a focused community . . . . .	121
5.2.1.2	Orderly heterogeneity and individual agency . . . . .	122
5.2.2	Methodological innovations . . . . .	123
5.3	Subjects, objects, and copulas in Vietnamese . . . . .	124
5.3.1	Subject pronominal forms in Vietnamese . . . . .	124
5.3.2	Object pronominal forms in Vietnamese . . . . .	127
5.3.3	Copulas in Vietnamese . . . . .	129
5.4	Previous studies on the realisation of subjects, objects, and copulas in a cross- generational context . . . . .	131
5.4.1	The transmission of subjects, objects, and copulas across generations .	131
5.4.1.1	Subjects . . . . .	131

5.4.1.2	Objects . . . . .	133
5.4.1.3	Copulas . . . . .	134
5.4.1.4	Summary . . . . .	134
5.4.2	Pragmatic norms and cultural distance in language contact . . . . .	135
5.5	Analysing CanVEC: Data coding and method . . . . .	137
5.5.1	Coding the dependent variables . . . . .	137
5.5.1.1	Subjects . . . . .	138
5.5.1.1.1	Exclusion . . . . .	138
5.5.1.1.2	Partial exclusion . . . . .	139
5.5.1.2	Objects . . . . .	142
5.5.1.2.1	Exclusion . . . . .	142
5.5.1.2.2	Partial exclusion . . . . .	144
5.5.1.3	Copulas . . . . .	144
5.5.1.4	Corpus distribution: CanVEC subjects, objects, and copulas across generations . . . . .	146
5.5.2	Coding the independent variables . . . . .	146
5.5.2.1	Subjects . . . . .	147
5.5.2.1.1	Person-Number . . . . .	147
5.5.2.1.2	Clause Type . . . . .	148
5.5.2.1.3	Coreferentiality . . . . .	149
5.5.2.2	Objects . . . . .	151
5.5.2.2.1	Coreferentiality for objects . . . . .	151
5.5.2.2.2	A note on Animacy . . . . .	153
5.5.2.3	Copulas . . . . .	154
5.5.2.3.1	Predicate Type . . . . .	154
5.5.2.3.2	Subject Type . . . . .	155
5.5.2.4	Extra-linguistic factors . . . . .	155
5.5.2.5	Summary . . . . .	156
5.5.3	Statistical modelling: Rbrul mixed-effects . . . . .	156
5.5.3.1	Rbrul explained . . . . .	156
5.5.3.2	Rbrul modelling . . . . .	158
5.6	Results . . . . .	159
5.7	Discussion: Heritage language in the community . . . . .	163
5.7.1	Cross-generational variation: Traces of community bricolage . . . . .	163
5.7.1.1	The peculiar direction of effects for null subjects . . . . .	163

5.7.1.2	Inter-speaker variability . . . . .	165
5.7.1.3	Has Vietnamese co-evolved with speakers' English? . . . . .	168
5.7.2	Beyond cross-generational variation: Stability of other conditioning factors . . . . .	169
5.7.2.1	Coreferentiality effects . . . . .	170
5.7.2.2	Different effects of environmental factors . . . . .	171
5.7.2.3	A note on overt forms . . . . .	171
5.8	Chapter summary . . . . .	172
<b>6</b>	<b>Probing interface vulnerability: on the (over)use of overt forms</b>	<b>175</b>
6.1	Introduction . . . . .	175
6.2	Background . . . . .	176
6.2.1	The over-extension of pragmatic contexts of overt forms . . . . .	176
6.2.2	The vulnerable nature of the interfaces . . . . .	180
6.2.3	Summary . . . . .	184
6.3	Analysing CanVEC overt forms . . . . .	185
6.3.1	Defining appropriateness . . . . .	185
6.3.2	Coding overt pronominal subjects . . . . .	186
6.3.2.1	Redundant overt pronominal subjects . . . . .	186
6.3.2.2	Type of pronominal subjects . . . . .	188
6.3.2.3	Form of pronominal subjects . . . . .	190
6.3.3	Coding overt pronominal objects . . . . .	191
6.3.3.1	Redundant overt pronominal objects . . . . .	191
6.3.3.2	Type of pronominal objects . . . . .	192
6.3.3.3	Form of pronominal objects . . . . .	193
6.3.4	Coding overt copulas followed by adjectival predicates . . . . .	194
6.4	Results . . . . .	195
6.4.1	Redundant overt subjects . . . . .	196
6.4.2	Pronominal type and pronominal form . . . . .	198
6.4.3	A note on overt objects and overt copulas . . . . .	200
6.5	Chapter summary . . . . .	202
<b>7</b>	<b>Concluding remarks</b>	<b>203</b>
7.1	Heritage Vietnamese in the Canberra bilingual community . . . . .	204
7.2	Heritage languages in a broader context . . . . .	206
7.3	Where to from here? . . . . .	207

<b>Appendices</b>	<b>211</b>
<b>A Invitation letters to participate</b>	<b>211</b>
A.1 Vietnamese version . . . . .	211
A.2 English version . . . . .	212
<b>B CanVEC scores on language attitude</b>	<b>213</b>
<b>C Information and Consent Form</b>	<b>215</b>
C.1 Vietnamese version . . . . .	215
C.2 English version . . . . .	218
<b>D CanVEC corpus constitution</b>	<b>221</b>
<b>E Questionnaire</b>	<b>223</b>
E.1 Vietnamese version . . . . .	223
E.2 English version . . . . .	228
<b>F CanVEC annotation conventions</b>	<b>233</b>
<b>G Vietnamese to Universal POS tag map</b>	<b>235</b>
<b>H Vietnamese POS-tag confusion matrices</b>	<b>237</b>
H.1 Sample Vietnamese clauses . . . . .	237
H.2 Sample mixed clauses . . . . .	238
<b>I CanVEC example of an annotated dialogue</b>	<b>239</b>
<b>J CanVEC Vietnamese intransitive verbs</b>	<b>245</b>
<b>K Vietnamese null subjects per speaker per grammatical person</b>	<b>247</b>
K.1 First-generation speakers . . . . .	248
K.2 Second-generation speakers . . . . .	249
<b>References</b>	<b>251</b>





## LIST OF FIGURES

---

2.1	Geographical distribution of the Vietnamese community in Australia . . . . .	12
2.2	CanVEC demographics vs. Canberra Vietnamese demographics . . . . .	25
2.3	CanVEC speakers' primary language of social network . . . . .	29
3.1	Praat demonstration of Vietnamese pitch reset at IU boundaries . . . . .	42
3.2	ELAN demonstration of CanVEC speech-tier alignment . . . . .	54
4.1	The 4M model morpheme types . . . . .	66
4.2	The Matrix Language Turnover Hypothesis . . . . .	70
4.3	Applicability of the MLF principles to CanVEC mixed clauses . . . . .	91
4.4	Matrix-language ambiguous clauses with congruent word order . . . . .	92
4.5	Matrix-language distribution across generations . . . . .	107
5.1	The distributional pattern of the first generation's Vietnamese null subjects . . .	166
5.2	The distributional pattern of the second generation's Vietnamese null subjects .	167
6.1	Core and non-core components in the Y-model . . . . .	181
6.2	Gender and age index of basic Vietnamese pronominal forms . . . . .	187



## LIST OF TABLES

---

2.1	CanVEC demographic information . . . . .	24
2.2	Speakers of CanVEC . . . . .	27
2.3	CanVEC cross-generational language attitude . . . . .	31
2.4	CanVEC self-reported behaviours and attitudes towards language mixing . . . . .	31
3.1	Step-wise automatic annotation of CanVEC . . . . .	50
3.2	CanVEC frequent and widespread single other-language items . . . . .	53
3.3	Basic linguistic statistics of CanVEC . . . . .	54
3.4	Accuracy report of CanVEC semi-automatic annotation . . . . .	55
3.5	Pronoun and classifier POS-tag error distribution . . . . .	56
3.6	An overview of the translation quality in CanVEC . . . . .	58
4.1	Vietnamese-English word order differences in the nominal domain . . . . .	88
4.2	An overview of CanVEC difficult data in relation to the MLF . . . . .	106
4.3	Contrastive distribution of ML-identifiable CP types across generations . . . . .	107
4.4	Vietnamese generic classifiers in CanVEC otherwise-English clauses . . . . .	109
5.1	Excluded contexts for Vietnamese pronominal subjects . . . . .	142
5.2	Excluded contexts for Vietnamese pronominal objects . . . . .	145
5.3	Excluded contexts for Vietnamese copulas . . . . .	145
5.4	Cross-generational distribution of null vs. overt subjects, objects, and copulas . . . . .	146
5.5	Coding scheme for Vietnamese person-number . . . . .	148
5.6	An overview of the coding scheme for subjects, objects and copulas . . . . .	157
5.7	Factors included in Rbrul modelling . . . . .	160
5.8	Rbrul results for Vietnamese null subjects . . . . .	162
5.9	Cross-generational distribution of null vs. overt subjects, objects, and copulas . . . . .	172
6.1	Interface association of overt subjects, objects, and copulas in heritage Vietnamese . . . . .	183
6.2	Pragmatic distinctions between different Vietnamese pronominal types . . . . .	186

---

6.3	Distribution of different pronominal types for overt subjects in CanVEC . . . .	188
6.4	An overview of different pronominal types for overt objects in CanVEC . . . .	192
6.5	Results of Vietnamese overt subjects, objects, and copulas in CanVEC . . . .	195
6.6	The distribution of inappropriate Vietnamese overt pronominal forms . . . .	199

## ABBREVIATIONS

---

1	First person
2	Second person
3	Third person
ACC	Accusative
ADJ	Adjective
ADV	Adverb
AGR	Agreement
ASP	Aspect
AUX	Auxiliary
CLS	Classifier
COMP	Complementiser
COP	Copula
DAT	Dative
DEM	Demonstrative
DET	Determiner
DM	Discourse marker
EL	Embedded Language
ERG	Ergative
EXPL	Expletive
F	Feminine
FUT	Future
IMP	Imperative
IMPERF	Imperfect
INFL	Inflection
INTJ	Interjection
INTSF	Intensifier
IU	Intonation unit
kin	Kin terms

L1	First language
L2	Second language
LOC	Locative
M	Masculine
ML	Matrix Language
MLF	Matrix Language Framework
MP	Minimalist Program
NEG	Negation
NOM	Nominative
O	Object
PASS	Passive
PERF	Perfect
PL	Plural
POSS	Possessive
PREP	Preposition
PROG	Progressive
PRON	Pronoun
PRT	Particle
PST	Past
Q	Question marker
REFL	Reflective
S	Subject
SG	Singular
T	Tense
TOP	Topic
V	Verb

# INTRODUCTION

---

## 1.1 Setting the stage

As multilingualism increasingly becomes the norm, the interaction between language and the movement of people has become a central focus in sociolinguistics. Work in the past fifty years or so has addressed questions concerning, among other things, code-switching (the alternation between two languages in a single discourse/utterance), bilingual language acquisition and retention, contact-induced change, and cross-generational language variation and shift. This dissertation situates itself in relation to the last strand, specifically focusing on Vietnamese as a heritage language in Canberra, Australia. The subjects of investigation are late bilingual immigrants, whom I refer to as first-generation speakers (Gen 1), and early bilinguals raised in Canberra, whom I refer to as second-generation speakers (Gen 2). The ultimate aim of this dissertation is to characterise aspects of the cross-generational variation of Vietnamese as the heritage language.

With this in mind, it is first important to clarify two key terms that are contentious but play a crucial role in the subsequent discussion of this work: **heritage language** and **heritage language speakers**. Early definitions in heritage linguistics often defined a heritage language as one that is ‘spoken by early bilinguals, simultaneous or sequential, whose home language (L1) is severely restricted because of insufficient input’ (Polinsky, 2011, p.1). As the field began to mature, this emphasis on insufficient input as a qualifying characteristic has noticeably decreased. In fact, although different researchers still have different definitions of what the ‘heritage’ component entails, most today agree that a heritage language is a complete system on its own, and that the multi-faceted circumstances in which the heritage language is operating can make a decisive difference to speakers’ linguistic behaviour (see e.g. Polinsky, 2018; Aalberse, Backus & Muysken, 2019 for a helpful overview). For immigrant early-bilingual heritage language speakers in partic-

ular, the sole focus on divergence from a monolingual baseline can be rather meaningless, given that the input for their heritage language acquisition may come from the late bilinguals who are themselves outside their monolingual milieu (Polinsky, 2018; Polinsky & Scontras, 2020). In this sense, a study of an immigrant heritage language is not just a study of early bilinguals per se, but is in fact an enquiry into the transition from Gen 1 to Gen 2 speakers.

In the context of this work, I take a broad view of heritage language as a sociolinguistic construct, which involves speakers' agency and identity work, as much as their acquisition and proficiency. A heritage language is thus defined as 'a culturally or ethnolinguistically minority language that develops in a bilingual setting where another socio-politically majority language is spoken' (Montrul, 2015, p.2).<sup>1</sup> Speakers of a heritage language are speakers who use the minority variety as part of their repertoire. Patterns of acquisition and proficiency in the heritage language are not defining characteristics, as long as speakers can participate in spontaneous speech with communicative intent. In this sense, both Gen 1 and Gen 2 speakers are considered Vietnamese 'heritage language speakers' in the present work.

Returning to the broader context, this research is primarily motivated by a lacuna in the current body of literature on language variation and change, where work on minority languages and on the communities where these languages are spoken is still rather limited, especially in comparison to English and other Indo-European languages. As Stanford (2016, p.528) highlights, this lack of linguistic and geographical diversity is scientifically problematic, as 'the farther we move from the traditionally studied communities, the more likely we will see fieldwork results that challenge existing notions and principles—or at least cause us to reconsider assumptions and view them in a new light.' Given that contact Vietnamese has only been sparsely considered (Tuc, 2003; Thai, 2005; Nguyen, 2018), and that the Canberra Vietnamese bilingual community has never been investigated, this dissertation sets out to offer new data that enables us to potentially reconsider assumptions of this kind.

The objectives of the research are therefore twofold:

- (i) to document the vernacular of the Canberra Vietnamese community; and
- (ii) to consider key aspects of the data in relation to cross-generational variation, from prominent theoretical perspectives (specifically, the Matrix Language Framework, the variation-

---

<sup>1</sup>This also coincides with the Australian Government's definitions, which differentiate 'community languages' (minority languages of migrants) from 'Aboriginal and Torres Strait Islander languages' (native Australian languages) (Australian Bureau of Statistics, 2017). Note that the term 'community language' is preferred over 'heritage language' in Australia as it does not imply any language loss, historical association or discursive resonance (Liddicoat, 2018, p.237). Despite this, however, I use the term 'heritage language' in this work to be consistent with the broader literature on this topic.

ist approach, and the generative interface vulnerability approach) on data of the relevant kind.

The dissertation therefore has both empirical and theoretical objectives.

## 1.2 Research components

In this section, I outline the data and the theoretical perspectives on which the dissertation centres.

### 1.2.1 Data: Introducing the Canberra Vietnamese-English Corpus

The first objective, to document the vernacular of the Canberra Vietnamese community, inspired the creation of the Canberra Vietnamese-English Corpus (CanVEC). The corpus was newly compiled for the present study and consists of over 10 hours of spontaneous speech produced by 45 Vietnamese-English bilingual speakers from two generations living in Canberra, Australia. The vernacular documented in the corpus features speakers' monolingual Vietnamese, monolingual English, as well as their code-switching production. Example (1), drawn from the corpus, illustrates this diversity in continuous speech as part of a natural dialogue. Every CanVEC example presented features a transcript name (e.g. Hannah.Lida.0718 in (1)) and a timestamp, with the subscript accompanying the speaker name indicating their generation membership (1 = Gen1; 2 = Gen2).<sup>2</sup> English is given in regular print, while all non-English morphemes are given in *italics* throughout this dissertation.

- (1) a. Hannah<sub>2</sub>: and then on Wednesday I have netball training, [Monolingual English]  
 b. *xong-rôi nó vậy thôi đó.* [Monolingual Vietnamese]  
 then EXPL like-that DM DM  
 'Then that's it.'  
 c. on Thursday I sometimes *đi nhà-thờ.* [Code-switching]  
 go church  
 'On Thursday sometimes I go to church.'
- (Hannah.Lida.0718, 12:46.6–12:57.9)

As example (1) illustrates, the corpus includes different varieties produced by the same speakers. While the primary focus of this work is on Vietnamese as the heritage language, the presence of English and code-switching discourse produced by the same speakers is extremely relevant. For example, as we will see in [Chapter 4](#), examining code-switching utterances may advance our understanding of the dynamics between the languages participating in this code-switching, i.e. the majority language (English) and the heritage language (Vietnamese). Similarly, direct com-

<sup>2</sup>A more detailed description of the transcript file labelling convention can be found in [Chapter 3](#), §3.3.1.1.

parison of the patterns in speakers' English and Vietnamese will also allow us to gauge the extent to which these languages interact and influence each other.

To maximise its future use, CanVEC is semi-automatically annotated with language marking, Part-of-Speech (POS) tags and translations. The corpus consists of approximately 90,000 words and 14,000 clauses, and is freely available for research purposes. The corpus serves as the basis for the analyses that follow.

## 1.2.2 Theoretical frameworks

The second objective, to characterise the cross-generational difference in the Vietnamese of the Canberra community from prominent theoretical perspectives, means that I do not rigidly subscribe to a single framework. Instead, three different frameworks are deployed at different stages in order to unpack different aspects of the empirical pattern: the Matrix Language Turnover Hypothesis (based on the Matrix Language Framework) (Myers-Scotton, 1993, 2002), the variationist framework (Labov, 1972 et seq.), and the generative interface vulnerability approach (Sorace & Filiaci, 2006; Sorace & Serratrice, 2009a; Sorace, 2011; Tsimpli, 2014; Sorace, 2016, i.a.). In what follows, I briefly introduce these models.

The Matrix Language Framework (MLF) approach ([Chapter 4](#)), first proposed by Myers-Scotton (1993), assumes an asymmetrical relationship between the two languages in a bilingual discourse. Specifically, the assumption is that speakers and hearers generally agree on which language the mixed sentence is coming from (Joshi, 1985, pp.190–191), and that this language constitutes the 'Matrix Language' (ML) of the conversation. In a code-switched clause, the MLF predicts that the ML:

- (i) supplies closed-class morphemes such as function words; and
- (ii) determines word order.

The Matrix Language Turnover Hypothesis then refers to a situation in which the original ML, i.e. the language that provides the morphosyntactic frame for a bilingual Complementiser Phrase (CP, which is roughly a clause), becomes the structurally Embedded Language (EL) in a given community and vice versa. In most cases, the original ML is the minority language (i.e. the language with less socio-political power), whereas the EL is the language of the majority (i.e. the language with more socio-political power). Due to higher prestige and/or greater socio-economic and political power, the majority language takes over and replaces the minority language as the ML for most bilingual CPs produced by community speakers. The Matrix Language Turnover Hypothesis then states that when a cross-generational 'ML Turnover' occurs, i.e. when the EL

in one generation becomes the ML in the other generation, language shift or language death will follow.<sup>3</sup> Studying a ‘ML Turnover’ is therefore potentially illuminating in capturing ongoing changes within the community and envisioning the most likely future of the heritage language.

Another motivation for adopting the MLF model and its associated ML Turnover Hypothesis is that although proponents of the model claim ‘universality of support, no matter which languages are involved’ (Myers-Scotton, 2006, p.248), support for this asymmetrical model has mainly come from language pairs that are typologically different in terms of their clausal word order, or else have vastly different inventories of inflectional morphology. Some examples include Myers-Scotton (1993) on Swahili-English (agglutinative-analytic); Fuller & Lehnert (2000) on German-English (fusional-analytic); Deuchar (2006), Deuchar, Davies & Donnelly (2018) on Welsh-English (VSO-SVO); and Wang (2007) on Tsou-Mandarin (VOS-SVO). A language pair such as Vietnamese-English, in which both languages are SVO and morphologically limited, has rarely been discussed (see Wang, 2007, 2016, however, for Mandarin-Southern Min). Data featured in CanVEC thus offers an enticing testing ground for the widely-held ML theoretical assumptions.

The variationist approach (Chapter 5), on the other hand, does not assume a ‘Matrix Language’ per se, but takes as central the regularity that underlies the variation of the languages as they are spoken within the community (Labov, 1972). The ultimate aim of the variationist approach is to reveal patterns and pinpoint sociolinguistic constraints that underlie a specific linguistic variable. Within this framework, the application of any grammatical rule is probabilistic rather than categorical, and the presence or absence of certain features makes the application of this rule more or less likely. As such, the conditions under which any preferred pattern of usage applies are weighted by quantitative statistics. The key advantage of the variationist approach is that it allows the heritage language to be examined as it is spoken in the community, without reference to any idealised benchmark. This not only holds significant descriptive value, but also allows us to identify trends and the direction in which the heritage language appears to be evolving.

The final framework that I adopt in this study (Chapter 6) is a generative framework which focuses on interface vulnerability in language contact (Sorace & Filiaci, 2006; Sorace & Serratrice, 2009a; Sorace, Serratrice, Filiaci & Baldo, 2009b; Sorace, 2011; Tsimpli, 2014; Sorace, 2016).

---

<sup>3</sup>Language shift refers to ‘the gradual displacement of one language by another in the lives of the community members’ (Dorian, 2014, p.205). It typically manifests as a majority language taking over a minority language in terms of use. Language death is one of the possible eventual outcomes in a very extreme situation of language contact, where a language ‘stops being used by a speech community while another language expands in all domains and is passed on to the next generation’ (Dal Negro, 2004, p.47). Therefore, with the exception of rare cases of ‘sudden death’ (i.e. a language dies because an entire speech community vanishes as a result of war, genocide or natural catastrophes, etc.), the most usual context of language death is one of bilingualism, or rather, of ‘a very unstable and asymmetrical kind of bilingualism in which two or more languages are in contact’ (ibid.).

This approach makes a distinction between ‘core’ and ‘non-core’ linguistic phenomena, which are respectively linked to early- and late-acquired properties. The core properties are those that belong to narrow syntax, while the non-core, late-acquired phenomena are those that involve the intersection of different language modules (e.g. phonology, lexicon, morphology, syntax, etc.). Non-core properties are more demanding in terms of linguistic and other cognitive resources, and thus expected to be more vulnerable under contact. In this study, the adoption of this approach is motivated by some key observations in [Chapter 4](#) and [Chapter 5](#), which neither the MLF nor the variationist model can account for.

Ultimately, the hope is that this integrated approach will serve to showcase the value of exploring different aspects of the data using multiple theoretical lenses, thereby contributing to the attempt to reconcile traditionally divergent voices in research on language contact and variation (Cornips & Corrigan, 2005, p.2).

### 1.3 Overview of the study

This dissertation consists of six further chapters. The following five chapters are the core chapters, and [Chapter 7](#) is the conclusion. The central five chapters are organised into two parts. In Part I (Chapters 2–3), I address the first goal of the study and describe the Canberra Vietnamese community and the construction of the Canberra Vietnamese-English Corpus (CanVEC). In Part II (Chapters 4–6), I put the corpus to use and address the second goal of the research, characterising aspects of the cross-generational differences in Vietnamese as the heritage language in this community.

Part I begins with [Chapter 2](#), which describes the contact settings of the community and the speakers involved in the study. Key to the discussion are the specifics of the community that set it apart from other typical Vietnamese migrant diaspora elsewhere in Australia, especially in terms of how there is neither a designated Vietnamese neighbourhood nor previous evidence for a well-established community ‘speech norm.’ I discuss how this challenges the traditional boundaries of a ‘speech community’ in sociolinguistics, and argue in favour of a combination of different indicators of how this group of speakers functions both as a ‘speech community’ (Labov, 1972) and as a ‘community of practice’ (Eckert & McConnell-Ginet, 1992). The chapter also provides the relevant demographic and linguistic information about the speakers in the corpus.

[Chapter 3](#) describes in more detail the construction of CanVEC. Central to this chapter is a description of the data collection, transcription and annotation method. An additional contribution to the field is a newly-developed toolkit—an outcome of collaborative work with the Cambridge Computer Laboratory—to semi-automate language identification, Part of Speech (POS)

tagging and translation in the corpus (Nguyen & Bryant, 2020). As this process is notoriously time-consuming and laborious to undertake manually, the development of this toolkit is an attempt to streamline the creation of similar low-resource language corpora in the future. Together, [Chapter 2](#) and [Chapter 3](#) contextualise the linguistic analyses in the second part of the dissertation.

Part II begins with [Chapter 4](#), where I investigate cross-generational language variation and shift in the community and put the ML Turnover Hypothesis to the test. Specifically, I probe the Vietnamese heritage language via its participation in the bilingual discourse and explore whether an ML Turnover is underway (or complete). At the heart of this chapter is the discussion of the applicability of the MLF principles to a language pair like Vietnamese-English, i.e. a pair with limited morphology and shared SVO clausal word order, as well as the predictive power of the ML Turnover Hypothesis. As this chapter will show, various aspects of the CanVEC data challenge the existing MLF assumptions.

In light of the shortcomings of the MLF and the ML Turnover Hypothesis that emerge in [Chapter 4](#), [Chapter 5](#) continues the enquiry by moving away from the MLF and the bilingual portion of CanVEC to examine the monolingual heritage Vietnamese subset directly. Having identified the problematic nature of referencing an idealised monolingual norm, this chapter uses the variationist approach as an alternative to circumvent this problem. As null elements emerge as a distinct area of difficulty in [Chapter 4](#), I take the distinction between null and overt realisation of functional elements as the focus of further investigation here. In this chapter, these null elements are probed via three cases where the null and overt alternation arises in Vietnamese: subjects, objects, and copulas.

While null forms are comprehensively investigated in [Chapter 5](#), it has frequently been suggested that the overt counterparts of null forms exhibit distinctive behaviour in bilingual contexts of different kinds. The overt counterparts of the null subjects, objects and copulas in [Chapter 5](#) are therefore the focus of [Chapter 6](#). Here, I appeal to the interface-oriented approach that has featured strongly in recent generative discussion, seeking to establish whether the different interface factors regulating the occurrence of overt subjects, objects and copulas in colloquial Vietnamese have been preserved in the Canberra community, or whether this community also exhibits interface vulnerability effects of the kind that have been uncovered in other bilingual communities. Although the variationist approach offers extensive descriptive values, this chapter will show that the interface-oriented approach brings the focus back to the underlying explanatory factors concerning what conditions the vulnerability of a given property in contact.

In [Chapter 7](#), I finally bring these findings together, discuss their implications, and highlight possibilities and priorities for future research.



## **Part I**

# **Documenting the Canberra Vietnamese community vernacular**



# CHARACTERISING THE COMMUNITY: VIETNAMESE IN CANBERRA

---

## 2.1 Introduction

The first objective of this work is to document and describe the vernacular in the Canberra Vietnamese bilingual community. The first step in doing so is to delineate its speakers and the social landscape of their languages. In this chapter, I thus discuss the general political history and language use of the community. I begin with the Vietnamese community in Australia generally (§2.2), before describing the Canberra Vietnamese community in particular (§2.3). Next, I introduce the speakers who participated in this study, their demographic information and social networks, as well as their linguistic attitudes and practices (§2.4).

## 2.2 Vietnamese in Australia: Political history and language use

The Vietnamese diaspora is spread across a number of countries outside Vietnam, with the United States hosting the largest population (more than two million people), followed by Cambodia (600,000), France (350,000) and Taiwan (200,000), among many others. The Vietnamese population in Australia is the fourth-largest in the world, and home to 294,279 Vietnamese according to the 2016 Census. Figure 2.1 illustrates the distribution of the Vietnamese diaspora across the Australian states and territories.

As can be seen from the map, over three-quarters of Vietnamese immigrants reside in New South Wales and Victoria, particularly Sydney and Melbourne, which are the largest, most populated cities in the country. The community in Canberra in the Australian Capital Territory (ACT)

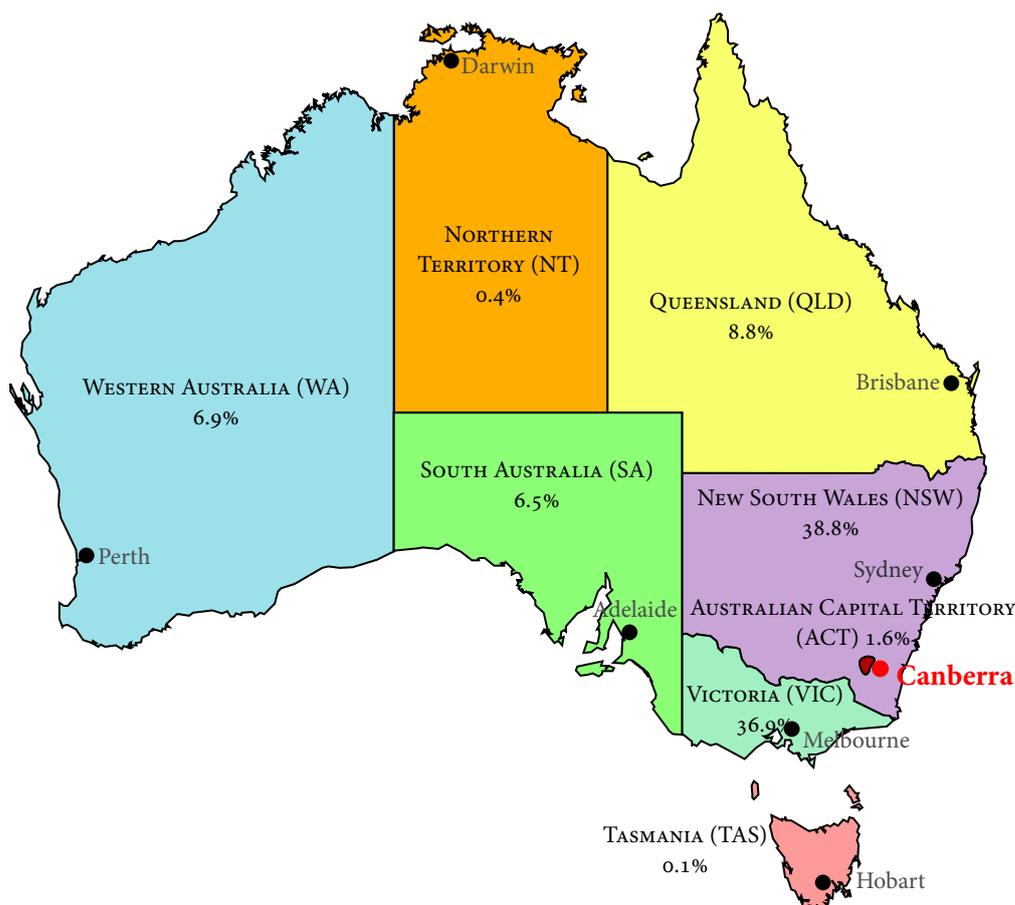


Figure 2.1: Geographical distribution of the Vietnamese community in Australia (N=294,279). Map source: <https://mapchart.net/australia.html>, data is added from the Australian Bureau of Statistics (ABS) Census 2016.

is relatively small in comparison, accounting for around 1.6% (N=4,216) of the Australian Vietnamese population overall.

A number of studies in language contact have shown that the linguistic landscape of a migrant community is likely to be affected by several distinct characteristics, most notably: the circumstances of arrival, the age at arrival and level of integration into the host society. Before the 1970s, there were only around 700 Vietnamese in Australia, most of them orphans adopted by Australian families, wives of Australian military personnel who had served in South Vietnam or tertiary students on a Colombo Plan scholarship.<sup>4</sup> Following the fall of Saigon in 1975, Australia received its first inflow of hundreds of thousands of Vietnamese arriving by boat. Vietnamese fleeing the war were recorded as arriving on an almost daily basis (Betts, 2001). This first wave mostly arrived as adults, with limited English literacy and few possessions. After receiving initial support for their predicament as refugees, the group became the target for traditional Anglo-Australian fears of an 'Asian invasion' (Carruthers, 2008a). This largely led to the con-

<sup>4</sup><https://www.destinationaustralia.gov.au/stories/work-play/colombo-plan>

gregation of Vietnamese in relatively highly-populated areas such as Springvale, Footscray and Richmond in Melbourne, or Bankstown, Marrickville and Cabramatta in Sydney. A report in 1986 showed that in Sydney alone, 30% of the 11,315 Vietnam-born men and almost 35% of the 7,496 Vietnam-born women in the working-age population were unemployed and looking for full time work (Burnley, 1989). This has been attributed to discrimination in the workforce, lack of functional English, and difficulties in having qualifications from Vietnam recognised.

Against this hostile political backdrop, the Vietnamese community rebuilt their life by setting up family businesses such as restaurants, grocery shops, manicure shops, hairdressing or cleaning services in the Vietnamese/Chinese-dense suburbs where they did not have to communicate in English on a regular basis. Their need to cluster has been recorded as a result of ‘on the one hand, experiences of racism and social exclusion in Australia, and on the other, the desire to be close to compatriots and to rebuild a sense of community’ (Carruthers, 2008a). It is thus no surprise that Vietnamese is particularly well-maintained in the community as a result. In a study of 466 Vietnamese speakers in 2012, Ben-Mosche & Pyke found that 90% of those surveyed reported being able to speak, read and write Vietnamese either ‘well’ or ‘very well’ (Ben-Mosche & Pyke, 2012). Many never learnt to speak English fluently. Vietnamese was also used widely within families, with more than 40% speaking to family members ‘always or mostly in Vietnamese,’ whereas just around 8% spoke to their children in English. The remainder of the group did not have children or spoke another language at home (Ben-Mosche & Pyke, 2012, p.31). Research has, however, found that many first-generation parents gradually adopted a large number of English words in their speech after their children started schooling in Australia. These typically include culturally loaded Australian terms such as ‘bungalow,’ ‘flat,’ ‘uni,’ etc. (Tuc, 2003).

After the continuous inflow of Vietnamese refugees from 1975 until the mid 1990s, a new wave of Vietnamese migrants began to arrive in Australia, primarily made up of international students, entrepreneurs and skilled workers. The proportion of refugees seeking asylum, as a result, also steadily declined over the years. Specifically, while refugees accounted for over 90% of Vietnamese migrants in the early 1980s, this number dropped to 22.7% in the early 1990s, and is now less than 1% of the total Vietnamese population (Australian Bureau of Statistics, 2017). This post-war second wave of immigrants arrived under very different circumstances to the first wave, bringing greater levels of education, more capital, and higher levels of English proficiency. The political attitude of this group also varied, depending on their background, their original region (North or South) and their association with the Communist Party. A sizeable number of international students in Australia in recent years are in fact funded by a Vietnamese Government scheme *Đề án 322* (Plan 322) or the Australia Awards.<sup>5</sup> These students often have some tie to the

<sup>5</sup><http://www.australiaawardsvietnam.org/index.php/en/about-us-2>

Communist Party in the homeland, such as working for a Government department, or being a member of the Party in general. Other new migrants arrive in pursuit of higher education, job opportunities in skilled occupations or investment in a state business as part of the economic visa scheme.<sup>6</sup> This subset of migrants in particular is often initially considered 'communist' and might face certain barriers as they integrate into the Vietnamese community.

It is also well-known within the community that while most identify as 'Vietnamese,' not many feel a strong connection towards the homeland. In Ben-Mosche & Pyke's (2012) study, they further found that despite an overwhelming proportion (88%) characterising themselves as Vietnamese, only just over half (51.5%) felt 'close' or 'very close' to Vietnam. A large minority (34%) expressed ambivalence towards Vietnam ('neither close nor distant'), whereas a small minority said that they felt 'distant' or 'very distant' overall. Such emotional distance is particularly prominent amongst the Australian-born (i.e. second-generation speakers), with 79% of those reporting that they felt 'distant,' 'very distant' or 'neither close nor distant' (p.30). This could be explained by the fact that while the first-generation speakers have a living memory of Vietnam, the second generation has grown up with little contact with the homeland. The report also states that as children of refugees in particular are likely to have been exposed to negative narratives about the Vietnam War, it is less easy for them to develop a positive sense of identity with the nation. In fact, most Vietnamese refugees in Australia are from South Vietnam and fled the war, and might still harbour considerable agony over the past. As Thomas (1999, p.185) describes, it is the realisation of:

- (i) how South Vietnam was transformed under the communist government;
- (ii) how the life they used to have ceased to exist; and
- (iii) the painful experience of risking their lives escaping the country

that has discouraged them from feeling attached to the homeland. These negative attitudes were most strongly manifested through three famous incidents:

- (i) when the former General Secretary of the Vietnamese Communist Party *Đỗ Mười* visited Canberra in 1996;
- (ii) when the Australian SBS news channel started screening VTV4, a TV channel run by the Vietnamese government for Vietnamese people living overseas in 2004; and
- (iii) when the Air-Vietnam-funded variety show *Duyên Dáng Việt Nam* arrived in Sydney for performances.

---

<sup>6</sup>See <https://immi.homeaffairs.gov.au/> for full details.

All these three events attracted unprecedented backlash from the Vietnamese community in Australia, first pressuring the then Opposition leader John Howard to not meet *Đỗ Mười*, and then causing the immediate cancellation of the VTV4 broadcast. At the variety show *Duyên Dáng Việt Nam*, thousands of protesters also picketed the performances, pronouncing this as ‘the latest and boldest initiative in an ongoing propaganda offensive aimed at infiltrating ‘communist’ popular culture into the overseas Vietnamese community’ (Carruthers, 2008b, p.72). Such political activities, among other things, have made the modern Vietnamese community highly visible to other Australians. While growing and diverse, the Vietnamese community continues as ‘distinctively Vietnamese’ (Ben-Mosche & Pyke, 2012, p.64).

The political tension has led to linguistic consequences. Most notably, the refugee group often distinguishes itself from ‘the communists’ living in the homeland by avoiding the use of terms adopted by the Communist Party after 1975. For example, ‘Ho Chi Minh City’ is often frowned upon in favour of ‘Saigon,’ *xe đò* is preferred over *xe khách* ‘bus/coach,’ or *tiểu bang* instead of *bang* ‘states.’<sup>7</sup> *Tiếng Việt Sài Gòn Cũ* (Old-Saigon Vietnamese variety) and *Tiếng Việt Cộng Sản* (Communist Vietnamese variety) are constantly juxtaposed and discussed within the Vietnamese community, once having been a topic of three consecutive sessions on the Vietnamese Radio Network in Australia (VNRA) (Nguyen, 2012, p.87). New migrants or those who do not have the same political backdrop thus often refrain from using words associated with *Tiếng Việt Cộng Sản* to avoid evoking hostility.

The Vietnamese language used in Australia is thus a combination of Old-Saigon Vietnamese, maintained by South Vietnamese, and modern Vietnamese homeland varieties from different sources, primarily new migrants. Vietnamese has appeared on the list of languages with the highest proportion of speakers in Australia in the bracket of 0-14 years old, together with Arabic, Lebanese, Khmer, Turkish, and Urdu (Kipp, Clyne & Pauwels, 1999). Nonetheless, it is still considered a low-status language in Australia, as it has never made the list of ‘high-priority languages’ for employment purposes.<sup>8</sup> Although Vietnamese has been introduced as a foreign language in some schools, it is often considered marginal in comparison with other Asian languages (Le, 1995, p.104).

In a more recent study on Vietnamese speakers in Australia, Nguyen (2015) found a correlation between parents’ level of education and language maintenance within the home. Specifically, the higher the parents’ level of education, the less their children spoke Vietnamese at home. Second-generation speakers with university-educated parents who participated in this study cited reasons such as ‘we do not need to speak Vietnamese,’ ‘my parents want to practise

<sup>7</sup>These lexical variants do not trigger any semantic differences. The choice of one form over another is purely a matter of preference.

<sup>8</sup>Six ‘high-priority’ languages in Australia include Chinese, Arabic, Japanese, Indonesian, German and French.

their English,' and 'everyone in my family speaks English fine' for their lack of engagement with the Vietnamese language. Speakers of parents without a degree, on the other hand, stated that because their parents are self-employed or unemployed, they usually 'cannot understand the Australian accent,' 'do not have functional English,' or 'speak no English at all.' As such, Vietnamese children from these families feel pressured to learn and practise Vietnamese regularly from a very young age, when they start realising their parents' limited level of English. Due to frequent usage, this group of second-generation speakers become fluent and enjoy speaking Vietnamese more (Nguyen, 2015, p.7). Most speakers in Nguyen's (2015) study also express 'a great deal of satisfaction' when they are able to switch back and forth between languages. Some speakers explain that it is harder for parents to learn English than for them to learn Vietnamese (due to old age and other factors), and so 'it just makes more sense if I try than forcing my parents to learn another language.' While intuitive, this finding is somewhat at odds with findings elsewhere, in which parents' proficiency and education level seemingly has no effect (Park & Sarka, 2007) or produces an effect in the opposite direction; that is, higher-educated parents are more likely to understand the value of the heritage language and subsequently make more effort to transmit it themselves to younger generations (cf. King & Fogle, 2006; Lee, 2012).

Against the backdrop of the complex political and linguistic background of the Vietnamese community in Australia, the next section characterises the Canberra Vietnamese community specifically, and considers how its defining characteristics diverge from the national landscape of the Vietnamese diaspora.

### **2.3 The Vietnamese community in Canberra**

The capital city of Australia, Canberra, is geographically located between Sydney and Melbourne, the two largest cities in the country. With a population of 406,403 spread over 814.2 km<sup>2</sup>, Canberra is the largest inland city of Australia, and the eighth largest city overall. One-third of Canberra residents are born overseas, with the most sizeable group coming from the United Kingdom (3.2%), followed by China (2.9%), India (2.6%) and Vietnam (1.2%) (Australian Bureau of Statistics, 2017). Although Canberra is still largely English-dominated (72.7% of locals speak only English at home), the latest 2016 Census shows that Vietnamese remains the second most popular heritage language spoken at home in the nation's capital (N=4,216), after Mandarin Chinese (N=12,408).

Of the Vietnamese community in Canberra, the majority are restaurant owners, students or workers in the public service (Australian Bureau of Statistics, 2017). Canberra residents in general are characteristically young, highly mobile and educated compared to the rest of the nation.

As of May 2017, just over one-tenth of the population in Canberra aged 65 and above, and almost half of those aged between 25-65, had achieved an educational level equal to a bachelor's degree. Canberrans are also the highest paid among the nation (average weekly income AUD 998, according to National Australia Bank, 2017). Such unified demographic features make the nation's capital an unusual, atypical social community. For the Vietnamese residents living in Canberra specifically, while official numbers are difficult to obtain, it is well-known in the community that this group fits into this overall picture and is typically 'Canberrans' for the most part: relatively young, well-paid and well-educated. Contrary to densely populated Sydney or Melbourne, in which Vietnamese speakers cluster in neighbourhoods and are employed in low-skill family business, the majority of Vietnamese speakers in Canberra work in education or the public sector, or have a partner doing so.

Against this backdrop, an important question to ask is whether this group of speakers forms a 'community' per se, and if so, how do we decide who belongs to the community and who does not. Admitting that a working definition of 'community' is difficult to formulate, Milroy (1980, p.14) notes two key factors as defining characteristics:

- (i) speakers' consciousness of belonging to a cohesive group; and
- (ii) the association to a strong territorial basis or 'localism.'

Milroy denotes 'localism' as typically constituted by a spatial concentration of the group members and the kinds of interactions that they engage in. Close-knit communities, she argues, derive most of their interactions from their neighbourhood, which in turn forms the heart of speakers' immediate 'social network.'<sup>9</sup> While this holds true for many communities, in what follows I will show how this definition conflates the concept of 'spatial concentration' with that of 'social network' for the Vietnamese community in Canberra (§2.3.1). I then make a case that, despite this difficulty with geographical delineation, there are still clear reasons to believe that close ties have been built for this group of Canberra Vietnamese speakers (§2.3.2).

### 2.3.1 Defining a 'speech community' for Vietnamese speakers in Canberra

Among various interpretations of what a 'community' might entail, the notion of 'speech community' has gone on to become one of the most influential in sociolinguistics. Specifically, the concept of a 'speech community' stems from so-called **first-wave** sociolinguistics, which focused on macro-sociological variables and the correlation between them and the use of different lin-

---

<sup>9</sup>It should be noted that in the context of this discussion, 'social network' is a theoretical construct that refers to a speaker's 'web of ties' and density of interactions within a community (Milroy, 1980) rather than the modern usage of internet social media. The precise influence of social media within a social network is beyond the scope of this study.

guistic features in a given community. Gumperz (1968) was one of the first to define ‘speech community,’ formulating it as ‘any human aggregate characterised by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language usage’ (p.381). Accordingly, two important components form the heart of a speech community: a set of linguistic norms and a set of social norms systematically shared among a group of speakers. One of the earliest and most well-known studies in this vein is Labov’s (1972) work on rhoticity in New York City. Labov showed a relationship between inter-speaker and intra-speaker variation in the production of post-vocalic /r/ as they were both connected to socio-economic class. Looking at employees working in three different department stores representing different socio-economic classes, he showed that the post-vocalic /r/ was earning prestige and spreading across New York City. This diffusion occurred at both the individual and the community levels, with the lower middle class responsible for the spread of the prestige form. Although the upper middle class was shown to use post-vocalic /r/ most often in casual registers, it was the lower-middle class speakers who led the change in formal speech. Labov took these findings as evidence for the existence of a New York City speech community, which exhibits both hierarchical differences and a shared set of norms. Labov then neatly summarised speech community as ‘not defined by any marked agreement in the use of language elements, so much as by participation in a set of shared norms; these norms may be observed in overt types of evaluative behavior, and by the uniformity of abstract patterns of variation which are invariant in respect to particular levels of usage’ (1972, p.120).

In this study, however, applying a Labovian definition of a speech community is not so straightforward. First of all, due to a serious lack of sociolinguistic work on heritage Vietnamese, evidence for ‘speech norms’ is limited at best, if not non-existent. The closest evidence we have to date is Nguyen’s (2018) study of single Vietnamese kin terms in an otherwise English context. There, I identified consistent and frequent use of Vietnamese kin terms in place of English pronouns for self- and interlocutor-reference. In the follow-up interviews, the speakers overtly rejected the English pronouns as viable alternatives, and the majority cited the community norm as a reason for this linguistic behaviour. This lends strong support for a unified speech norm, i.e. the shared dimension that is ‘related to ways in which members of the group use, value, or interpret language’ (Saville-Troike, 2003, p.15). However, my sample in that study was limited in size (15 speakers, 3 hours of recorded conversations), constrained to family conversations only, and consisted of speakers not specific to Canberra but from various Vietnamese communities in Australia. Evidence for a ‘Canberra Vietnamese speech norm’ therefore remains to be investigated.

In practice, sociolinguists often use geographical boundaries to delimit a speech community; however, such a focus on residence is lacking in any theoretical definition (Kiesling, 2011, pp.32–

33). As Kiesling states, while this is a convenient way to delimit the focus, it is problematic for several reasons. Specifically, while **place** can be defined by precise coordinates and boundaries, **space**—how people think about their physical surrounding—is not as straightforward. As an example, Kiesling discussed how people in Sydney tend to think of the city as a much more intermediate parameter, with areas like the ‘Northern Beaches’ having their own characteristics in talk about place. In other words, a theoretical distinction needs to be made between these two concepts, with **place** relating to physical coordinates and **space** being more of an interpreted reality. This distinction, in fact, is even more pronounced for a migrant diaspora, where speakers all have different histories of movement and different kinds of ‘ties’ to the community. For example, Mia (pseudonym), a speaker in CanVEC (Table 2.2), currently resides in Sydney, but goes back to Canberra every weekend to see her family and friends. Mia does not know any Vietnamese speakers in Sydney where she now lives, and still considers the Canberra Vietnamese network a big part of her identity and social circle. On the other hand, there are Vietnamese living in Canberra who do not consider themselves part of the Vietnamese community. For instance, Trung (pseudonym), a potential participant, contacted me upon reading my recruitment notice to tell me that my research was ‘full of unrealistic expectations.’ Accordingly, he advised that my best bet would be ‘to watch online Vietnamese shows like Paris By Night or Asia Got Talents (sic)’ for ‘realistic Vietnamese,’ as he did not think Vietnamese existed in Canberra any longer. As Table 3.3 in Chapter 3 will indicate, however, this position is not accurate as Vietnamese overwhelmingly remains the main medium of communication for the majority of the speakers in this study. What this incident shows is thus that there are some speakers like Trung, who, despite living in Canberra for 30 years, are still far removed from the Canberra Vietnamese network. This again highlights the danger of placing all speakers under a single geographic umbrella.

At this point, it seems clear that any attempt to operationalise a definition of a speech community is likely to be imperfect. However, as Kiesling (2011, p.33) points out, the important question that remains then is how do we decide who should be included in our sample, before beginning our data collection process? If geographical delineation seems insufficient and evidence of a shared speech norm is lacking, we need further information to ascertain the extent to which norms are shared and how often speakers talk to each other.

### 2.3.2 Canberra Vietnamese as a community of practice

It is now appropriate to turn to an alternative concept, which has become known as ‘community of practice.’ This was first introduced by Penelope Eckert in 1992. The idea is adopted in the so-called **second-** and **third-wave** sociolinguistic research, marking a transition of interest from the correlation between macro-social variables and linguistic features to how speakers use them to

construct identity. As Romaine (2012, p.446) puts it, ‘we all belong to many communities and sub-communities, defined in terms such as social class, ethnicity, nationality and religion.’ This view offers a much broader scope than the speech community definition. According to Romaine, this definition accounts for the multiple communities that we can be a part of, in which speech norms constitute only one of the components. Speakers of multiple bilingual communities around the world, such as Little Haiti in Miami, Little Italy in Boston, or Chinatowns in various places, for example, do not belong only to the communities of their heritage languages, but also to the global community of English speakers. What is key, for Romaine, is ‘the sense of perceived solidarity and interaction based on reference to a particular language and the relationships among people who identify themselves as members of that community’ (2012, p.447). In this sense, they form a ‘community of practice.’

In the case of the Vietnamese in Canberra, a strong indicator of an existing ‘community of practice’ is the maintenance of a group around shared activities, in which network ties are tightly construed (Kiesling, 2005). Specifically, the community hosts numerous activities such as regular charity stalls, weekly choir practice, karaoke nights, variety performances and *lễ phát phần thưởng* ‘end-of-year award-giving ceremonies’ as per the Vietnamese tradition, all of which contribute to the creation and maintenance of the group. The community also engage in regular interactions with each other at the Vietnamese language school in Dickson, North Canberra, where most families send their children for language classes every Saturday during term time. The engagement here is not only restricted to those with children, but also open to teachers and administrative staff at the school who are most often international students or retired members of the community. Parents and other members are often asked to help out with cooking and setting up for school events, where they also get to catch up with community news and with each other. Existing members usually bring along other Vietnamese speakers in their social network, and introduce them to other members of the school. Many have in fact been able to form new friendships and extend their own social network through these school-related cultural events. As will be discussed in §2.4.3, speakers have established a strong ‘exchange network’ for practical and emotional support, such as running errands, charities, dinner parties and so forth. The choice of the traditional dress *áo dài* for important cultural events such as Lunar New Year, weddings, and the Moon Festival is also another obvious marker of group membership (cf. Milroy, 1980). These shared practices are of particular importance in driving people closer to or further apart from a group identity, and have been recognised in both ‘community of practice’ analysis (Eckert & McConnell-Ginet, 1992; Wenger, 1998) and social network analysis (e.g. Milroy, 1980; Lippi-Green, 1989).

On this basis, the Vietnamese community in this study is defined as a group of speakers who share:

- (i) ethnicity as Vietnamese;
- (ii) a base in Canberra (though this is not always strict, as in the case of Mia);
- (iii) the languages they speak, i.e. Vietnamese and English;
- (iv) cultural events and practices; and
- (v) an exchange network in the community.

In fact, despite the defining characteristics of a ‘speech community’ being less straightforwardly evident, many speakers share all of the above. This definition also allows us to transcend the limits of geographical space entrenched in the traditional operationalisation of a ‘speech community’, thereby further narrowing down the circle of suitable speakers to identify those best representative of the community. For instance, speakers like Mia in the previous example would be included despite now living out of town, whereas Trung, who has almost no connection to the Canberra social network, is excluded from data collection.

### 2.3.3 Summary

To recap, this section described the Vietnamese community in Canberra, as well as how its characteristics diverge from other Vietnamese diasporas. Overall, the community comprises mostly relatively young, well-educated and well-paid speakers, who are often employed in highly-skilled jobs. The community is also unique in that it does not have a typical well-established clustered neighbourhood, with speakers spread across various parts of the city. However, despite these atypical demographics, I made a case that this group of speakers still effectively functions as a cohesive community (Eckert & McConnell-Ginet, 1992; Wenger, 1998), evident through their shared ethnicity, location, linguistic repertoire, strong exchange network and regular cultural practices.

On this principle, the next task is to assemble a speaker sample. This is a crucial step, as the credibility of the findings in any empirically-driven study is only as good as the data source. According to Torres Cacoullos & Travis (2018, p.35), specifying a linguistic data source ‘includes both the speaker sample and the observation method’: who are the bilinguals, and how was the data obtained? The following section addresses precisely these matters. A more technical aspect of the data collection component, however, forms part of a separate discussion, which I will address in [Chapter 3, §3.2.1](#).

## 2.4 CanVEC speakers: Who are they?

In this section, I introduce the speakers participating in this study, specifically the method of sample selection (§2.4.1), speakers' demographic profile (§2.4.2), social network (§2.4.3), and their language use and attitude (§2.4.4).

### 2.4.1 Pooling the sample

Between June and September 2017, I collected over 10 hours of spontaneous, informal speech produced by 45 Vietnamese-English bilingual speakers in Canberra and its surrounding regions within the Australian Capital Territory (ACT). This was a region where I had existing contacts with the community, having lived there for almost a decade, studying and working in Canberra. As Labov (1972, pp.114–115) recognised, the researcher's membership of the community offers an important advantage in community-based investigations, as established trust and networks with the speech community enable greater access to natural speech. This is even more true in the context of the ongoing political tension within the community, where Vietnamese living overseas remain sceptical of the so-called 'domestic Vietnamese' (§2.3.1) and are extremely wary of providing their data to 'the communists.' The fact that I left Vietnam many years ago and lived in Australia on a permanent basis was thus key in recruiting speakers and collecting quality data. With the help of existing family members in Canberra, my (re)integration into the community occurred quite smoothly and naturally. Participants were sought in two ways: from my informal contacts within the Canberra Vietnamese community, and via advertisements.

A fond memory of the fieldwork that further highlights speakers' strong sense of group membership was my struggle to persuade speakers to accept payment for their participation. Thanks to the availability of fieldwork funding, I was able to offer each speaker 40 Australian dollars for the recording and an additional 10 dollars for the completion of the questionnaire. However, as is typical of Vietnamese culture, many considered participation a favour for a community member rather than a paid task. One speaker said it was too 'Western' of me to do so, and that he was slightly disappointed that I assumed that his help *có thể quy ra tiền* 'could be measured by money.' This came not so much as a surprise knowing the Vietnamese culture. However, as there were other speakers who did not know me beforehand and had no difficulty accepting payment, I needed to ensure consistent ethical practice. This presented as much of a psychological challenge as a practical one, since I had to explain the research procedure without making myself appear too culturally distant. Second-generation participants were of great help in this regard, as they on many occasions explained to their parents/relatives (i.e. the other speaker) that it was only

fair if they accepted the payment for their contribution, and that it did not make them a ‘bad’ or a ‘greedy’ Vietnamese person in any way.

Additionally, I also used informally worded advertisements in both English and Vietnamese, including invitation letters ([Appendix A](#)) placed on bulletin boards and on several online platforms: the Language Diversity at ANU Facebook page<sup>10</sup>, the ANU Vietnamese Students’ Association<sup>11</sup> and Alliance, an internal linguistic forum at the Australian National University, to extend the range of speakers. Alliance was particularly helpful as it incidentally put me in touch with two students, Michael Carne and Li-Chen Yeh, who were doing phonetic research on Vietnamese at the time. They individually contacted me and kindly offered to introduce me to some of their own participants, a few of whom went on to become speakers in CanVEC.

### 2.4.2 Demographic profile and generation membership

As already indicated, the participants in CanVEC are the first- and second-generation speakers in the Canberra Vietnamese community. First-generation speakers are the first of their family to emigrate to Australia, lived in Vietnam at least till the age of 18, and have been living continuously in Canberra for at least 10 years. The first-generation speakers in the sample range in age from 28 to 67 as the sample consists of speakers belonging to several waves of immigration: some are refugees who fled the war more than 40 years ago, and some are recent economic migrants. Second-generation Vietnamese migrants are those whose parents qualify as first-generation speakers (even though they might not be in the corpus), and were either born in Australia or arrived together with their parents before the age of five. This benchmark of five years old for the second-generation was set to ensure not only that second-generation participants had been exposed to English-speaking communities prior to beginning school, but also that in the case of their arriving in Australia after birth, the amount of time that they had spent in their country of origin was minimal (cf. Kiesling, 2005, p.6; Hoffman & Walker, 2010, p.44).

It is thus important to stress that generation membership is not necessarily age-correlated in the context of this study. As [Table 2.1](#) illustrates, the youngest speaker in the first-generation bracket (aged 28) is younger than the oldest speaker in the second-generation (aged 35). The decision to not group younger speakers together as ‘Gen 2’ is justified on the basis that, both culturally and linguistically, migrants arriving as adults (refugees or economic) have more in

---

<sup>10</sup><https://www.facebook.com/groups/languageatANU/>

<sup>11</sup><https://www.facebook.com/ANUVietnamStudents>

	Speakers	Age Range	Gender		Education Level	
			Male	Female	University	School
<b>Gen 1</b>	28	28–67	15	13	19	9
<b>Gen 2</b>	17	12–35	6	11	9	8
<b>TOTAL</b>	45	12–67	21	24	28	17

Table 2.1: CanVEC demographic information

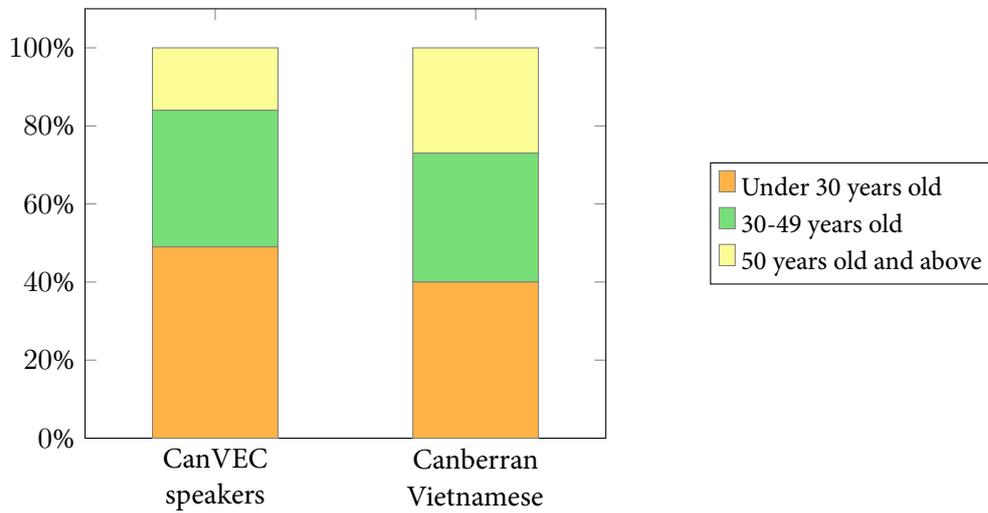
common with each other than with those born in Australia or who arrived as a young child.<sup>12</sup> It should also be noted that although the age range within the first-generation group of speakers seems rather large, finer-grained grouping (such as refugees vs. economic migrants) is not justified by the sample size and data distribution. This should not be a concern in this study, however, as we will later see that the grammatical patterns found for each generation are in fact strikingly consistent (while being distinct from each other, [Chapter 5](#)). This further reaffirms that the ‘right’ speakers were put into the ‘right’ group, at least for the variables that we are probing.

It is also clear from [Table 2.1](#) that over half of the speakers are thirty years and below (N=24, 53%). As we will also see from [Table 2.2](#), most speakers have pursued highly-skilled jobs such as engineer, lawyer, scientist, pharmacist, lecturer and the like. This distribution is obviously not representative of the wider Vietnamese population in Australia, but is a relatively accurate reflection of the demographics of the Vietnamese population in Canberra overall. [Figures 2.2a](#) and [2.2b](#) exemplify this distribution.<sup>13</sup> Benchmark data for the Canberra Vietnamese population, as shown in these figures, is drawn from the latest figures available, Census 2016. This was only one year before the CanVEC sample was collected, thereby making the data maximally comparable.

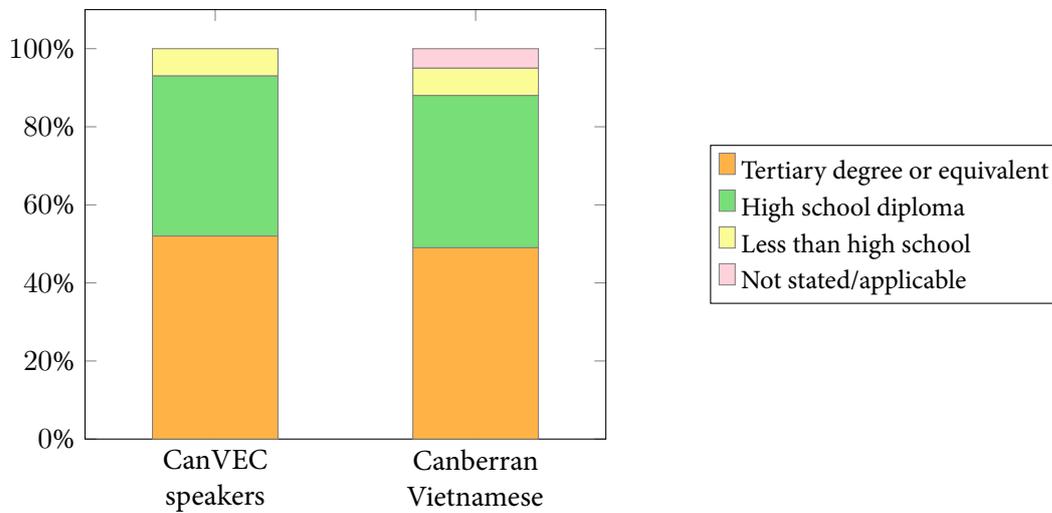
Finally, it should also be noted that although the Vietnamese language school (described in [§2.3.2](#)) serves to bring members of the community physically and socially together, none of the speakers in CanVEC were learners at the school. The school was mainly where most families sent their **young** children (under 10, who were not included in the corpus) to learn how to read and write in Vietnamese, through which they met and socialised on a regular basis. This means that while the language school is a defining aspect of the community, prescriptive teaching of Vietnamese should not be a concern for the patterns of language use acquired by CanVEC speakers.

<sup>12</sup>Although best efforts were made to recruit an equal number of first- and second-generation speakers, the nature of the data collection method and its emphasis on natural speech (detailed in [Chapter 3](#), [§3.2.1](#)) meant that it was very difficult to perfectly balance the dataset. It should be noted, however, that despite the difference in number of speakers across generations (28 vs. 17), each generation still produced a significant number of relevant clauses/tokens for each area of interest. In other words, there is no shortage of data for the results of any group, as we will see in [Chapters 4–6](#).

<sup>13</sup>Speakers younger than 15 at the time of the recording (N=5/45 for CanVEC and N=699/4,216 for the Canberra Vietnamese population) were excluded from the count for highest qualifications achieved, as they were still within the age range of minimum compulsory education in Australia.



(a) Age distribution of CanVEC speakers vs. the overall Canberra Vietnamese population



(b) CanVEC’s formal qualifications vs. the overall Canberra Vietnamese population (>15 years old)

Figure 2.2: Comparison of the distributional patterns of CanVEC demographics (including age and formal qualifications) with Canberra Vietnamese demographics, ABS Census 2016

Recording No.	Pseudonym	Generation	Gender	Year of birth	Occupation
01	Tee	1	M	1976	Engineer
01	Taz	1	F	1979	Lawyer
02, 22	Tim	1	M	1976	Scientist
02, 22	Jess	2	F	2002	Student
03	Mia	1	F	1985	Pharmacist
03	Phoebe	1	F	1982	Officer
04	Tanner	1	M	1976	Engineer
04	Nina	2	F	2003	Student
05	Theresa	1	F	1953	Dress-maker
05	Twee	2	F	1981	Graphic designer
06	Luna	1	F	1955	Shop owner
06, 07	Tressie	2	F	1991	Architect
07	Harry	1	M	1959	Builder
07	Josh	2	M	1992	Dancer
08	Mina	1	F	1978	Business manager
08	Pete	2	M	2004	Student
09, 12	Dany	1	F	1988	Counsellor
09	Lami	2	F	1990	Tutor
10	Helen	1	F	1974	Public servant
10	Vivian	2	F	2002	Student
10	Quinn	2	M	2003	Student
11	Marie	1	F	1953	Shop owner
11	Rory	1	M	1959	Restaurant owner
11	Penny	2	F	1987	Public servant
12	Brian	1	M	1989	Student
13	Lina	1	F	1978	Craft artist
13	Naomi	2	F	2001	Student
14, 23	Tom	1	M	1986	Kitchen hand
14, 23	Henry	2	M	1992	Cook

15	Quentin	1	M	1985	Student
15	Sony	1	M	1988	Gamer
16	Thomas	1	M	1984	Student
16	Max	1	M	1986	Student
17	Heather	1	F	1954	School teacher
17	Troy	2	M	1983	Waiter
18	Billy	1	M	1989	Unemployed
18	Ellie	1	F	1988	Pianist
18	Tyler	2	M	1987	Unknown
19	Quintus	1	M	1988	Student
19	Daniel	1	M	1987	Student
20	Reece	1	M	1949	Lifestyle assistant
20	Taylor	2	F	1988	Lecturer
21	Hannah	2	F	2002	Student
21	Lida	2	F	2001	Student
22	Chloe	1	F	1978	Accountant

Table 2.2: CanVEC speakers' demographics. The second number in the first column indicates the second recording in which the speaker participated.

### 2.4.3 Social network

Social network is an important extra-linguistic factor affecting patterns of language use (Milroy, 1980), and the correlation between these two variables has been demonstrated in several works (e.g. Milroy, 1980; Milroy & Wei, 1995; Tagliamonte, 2012). In order to understand the structure of networks within which the Canberra Vietnamese community interact, the personal social circles of each speaker were probed both via the questionnaire and via systematic content analysis of the recordings (see Torres Cacoullous & Travis, 2018). As Torres Cacoullous & Travis (2018) point out, the content—what people actually say in their own words—‘brings to the analyst’s attention issues and attitudes relevant to the community beyond the predetermined categories imposed by the questionnaire’ (p.62). In their study on the Spanish-speaking community in New Mexico, for example, not all categories emerging from the content correspond to items in the questionnaire, e.g. the stigmatisation of New Mexican Spanish, contact with Mexicans, or language choice in younger and older generations. Responses to questionnaires, furthermore,

can also be interpreted within the context of the community, contextualising, supplementing or even clarifying information obtained from the recordings.

Contrary to the general assumption of ‘weak ties’ associated with a geographically dispersed community, the lower density of Vietnamese migrants in Canberra in fact prompted speakers to be more proactive in seeking out other community members, which has in turn enabled them to establish a longstanding, strong ‘exchange network’ (Milroy & Wei, 1995). Speakers often rely on this network for practical support, such as babysitting, running errands, or sending money back to their families in Vietnam. For example, in the recording Heather.Troy.0506, the speakers were considering offering another speaker (who was also a participant, but in a different recording) their unwanted furniture; or in Quintus.Daniel.0806, Quintus was telling Daniel that he had been picked up from the airport by another member of the community whom he had not even met before arrival.

To investigate speakers’ social network, questionnaires are often used to ask participants to identify people that they have spoken to over the course of a unit of time (Gal, 1978, 1979). In Gal’s (1979) study of language shift in Oberwart (Austria), for example, this unit of time was defined as seven days; i.e. speakers were required to name all those they had spoken to in the past week. While this cut-off provides a useful window into speakers’ recent language use, it is intrinsically circumstantial and might not accurately reflect the language of the speakers’ regular social network. For some speakers, it might also be perceived as an infringement of privacy. In this study, I thus opted to ask participants to list the five people that they speak to most on a regular basis, their relationship, as well as the language they use. Following Deuchar et al. (2018), responses were given a numerical score, indicating whether they spoke English (1), Vietnamese (2), or both (3) with each contact. An average score was then calculated, and results are displayed in [Figure 2.3](#).

As [Figure 2.3](#) illustrates, more than half of CanVEC speakers (N=45) spoke both languages equally with their closest contacts on a daily basis. Over a third reported speaking mainly Vietnamese, whereas the remaining 16% spoke mainly English. An important qualitative result to report, which is not shown in the graph, is that all speakers named family members and other Vietnamese friends among their five closest contacts. This level of cohesion is consistent with what was reported for the Melbourne Vietnamese community (Tuc, 2003), a community with a higher population density and better-defined geographical concentration. The argument is thus strengthened for the existence of the Canberra Vietnamese ‘community,’ despite its lack of a clear-cut, well-established geographical concentration. As Tuc (2003, p.30) summarises, the high level of connection found reflects ‘the tradition of collective life-styles in Vietnam,’ or in other words, the cultural heritage values of the speakers.

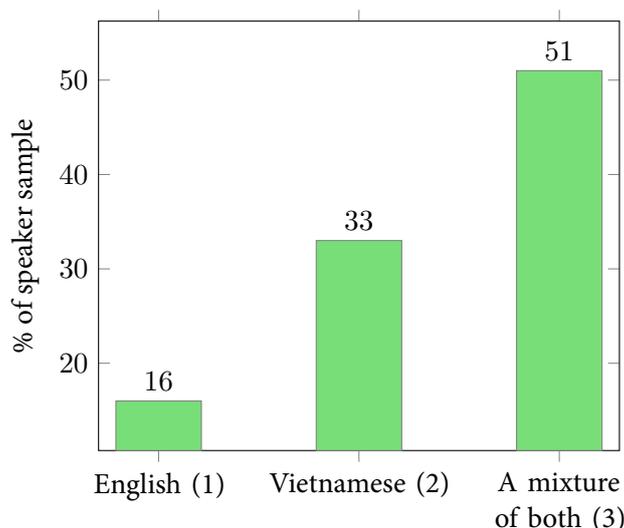


Figure 2.3: An overview of CanVEC speakers' primary language of social network

#### 2.4.4 Language maintenance, language attitude and language preference

All participants are literate in both languages, though to varying degrees of proficiency. In the aggregate, most CanVEC speakers rated themselves as, at the very least, 'fairly confident in extended conversations' in both English (N=41/45) and Vietnamese (N=36/45).<sup>14</sup> This sets the first-generation speakers of CanVEC in particular apart from first-generation Vietnamese speakers elsewhere, who are often found to have limited functional English (cf. Tuc, 2003; Ben-Mosche & Pyke, 2012). While a correlation has previously been drawn between 'the first generation's low level of English' and the 'dominance of Vietnamese for communication between all family members' (Ben-Mosche & Pyke, 2012, p.31), this is not necessarily true for the Vietnamese community in Canberra. Despite the generally high level of English, Vietnamese is still prevalent in the family domain (verified by the high proportion of Vietnamese utterances in the recordings, later shown in [Chapter 3, §3.3](#)). On two specific occasions, second-generation speakers were even explicitly asked to speak Vietnamese, as demonstrated in the following excerpts.

- (2) a. Pete<sub>2</sub>: could you trade my hundred over there for five?  
 b. Mina<sub>1</sub>: *con nói tiếng Việt đi.*  
 2SG.kin speak language Vietnam IMP  
 'Speak Vietnamese.'  
 c. *con muốn cái gì?*  
 2SG.kin want CLS what  
 'What do you want?'

(Mina.Pete.0906, 01:53.5–02:01.8)

<sup>14</sup>The full set of questions and answers is shown in [Appendix E](#).

- (3) a. Nina<sub>2</sub>: my friend Rita she was on,  
 b. Tanner<sub>1</sub>: can you speak Vietnamese?  
 c.           *con        nói với Rita sao?*  
               2SG.kin say with Rita how  
               ‘What did you say to Rita?’

(Tanner.Nina.0609, 14:28.0–14:47.1)

It should be noted, however, that such explicit requests to speak Vietnamese were only directed at the two youngest second-generation speakers in the corpus (aged 12 and 13 at the time of the recording), and are not found in any other recordings. This is to say that while these examples serve to illustrate the effort to maintain Vietnamese in the family domain, they cannot explain the proportion of Vietnamese spoken by the second generation.

I also measured language attitude using a questionnaire, which was modelled on one designed by Deuchar et al. (2018) to build their bilingual Welsh-English corpus *Siarad* (Chapter 3, §3.2.2). Specifically, speakers were asked to rate Vietnamese and English with reference to four pairs of adjectives, on a scale from 1 to 5, with 1 being the least and 5 the most positive. Three pairs described the general ‘feel’ of the language (‘friendly/unfriendly’, ‘inspiring/uninspiring’, ‘beautiful/ugly’), and only one pertained to the practicality of the language (‘useful/useless’). This question was designed to probe two types of language attitude: instrumental and affective (Garrett, 2010). Accordingly, speakers’ responses in relation to the ‘usefulness’ in the questionnaire represented their ‘instrumental attitudes’ to Vietnamese and English, whereas their reactions to ‘friendly’, ‘inspiring’ and ‘beautiful’ reflected ‘affective attitudes’, or feelings and emotions attached to the language.<sup>15</sup> Mean scores for each type of attitude in each generation in both languages were computed.<sup>16</sup> As Table 2.3 illustrates, English typically scores higher than Vietnamese both on instrumental and affective aspects.

<sup>15</sup>In the original questionnaire by Deuchar et al. (2018), two other pairs of adjectives—‘modern/old-fashioned’ and ‘influential/uninfluential’—were also used to measure instrumentality, together with ‘useful/useless’. However, these pairs were left out in this study due to the confusion they caused in the pilot run of the questionnaire. The descriptions of ‘modern/old-fashioned’ are not really applicable in the context of Vietnamese-English, given the lack of historical contact between the pair. Additionally, ‘influential/uninfluential’ also created some difficulties for pilot participants. They commented that while it was quite straightforward to rate the other three pairs of adjectives in relation to Vietnamese or English, the ‘influence’ of a language is more context-based: influential in terms of what, to whom, to what extent, and so on. Taking this feedback into consideration, I decided to omit this adjectival pair.

<sup>16</sup>In order to calculate the mean scores for each generation in each category, I first calculated the (mean) scores for each category per speaker. In particular, scores for ‘instrumental’ attitudes were recorded by the raw mark given by the participant to the adjective pair ‘useful/useless’, e.g. if a participant scored 2 for the level of usefulness of Vietnamese, then their ‘instrumental attitude’ in Vietnamese score was 2. On the other hand, scores for ‘affective’ attitudes are calculated as [the sum of individual scores for each ‘affective’ adjective pair/ 3], as the assessment of this category involves individual judgements of three separate pairs of adjectives. For example, if a participant scored 3 for Vietnamese being ‘friendly’, 4 for ‘inspiring’, and 1 for ‘beautiful’, then their affective attitude score for Vietnamese was  $[(3+4+1)/3] = 2.6$ . After each speaker was given a score for each type of attitude, scores for each category were totalled and averaged out by numbers of speakers in each generation. A more detailed record of speakers’ responses to each pair of adjectives can be seen in Appendix B.

	Instrument				Affective			
	Vietnamese		English		Vietnamese		English	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
<b>Gen 1</b>	<b>3.3</b>	0.3	<b>3.4</b>	0.6	<b>2.8</b>	0.5	<b>2.5</b>	0.7
<b>Gen 2</b>	<b>3.0</b>	1.0	<b>3.2</b>	0.6	<b>1.8</b>	0.3	<b>2.4</b>	0.8

Table 2.3: The distribution of language attitude scores across generations in CanVEC

The affective mean scores for Vietnamese stand out as rather low, particularly for second-generation speakers ( $\bar{x} = 1.8$ ,  $s = 0.3$ ).<sup>17</sup> This nonetheless somewhat fits in with the nation-wide trend previously reported in Ben-Mosche & Pyke's (2012) survey, in which they found that second-generation Vietnamese speakers exhibit a certain level of emotional distance from their motherland. While the raw difference in the mean scores seems significant at least in the affective aspect, a Welch's t-test found no statistical significance, either between generations or between the languages ( $p < 0.05$ ). This suggests that speakers in the community remain neutral about Vietnamese and English respectively, and in general do not see a great deal of disparity between the languages, instrumentally or affectively.

Speakers' reported views about language use and behaviour in the community, however, are not as unanimous. Participants' responses to the statement 'In everyday conversations I keep Vietnamese and English separate' and 'People should avoid mixing Vietnamese and English in the same conversation' in the questionnaire are particularly illuminating in this regard. As Table 2.4 shows, while a major proportion of speakers often hover around the middle ground, there is an almost even split on both sides of the spectrum (Agree or Disagree), particularly among the first-generation speakers. A clear majority of second-generation speakers, however, seem to be ambivalent about language boundaries, be it in their own behaviour (65%) or their general belief (41%).

Response	'I keep languages separate'		'People should keep languages separate'	
	Gen 1	Gen 2	Gen 1	Gen 2
<b>Total Disagree</b>	36% (N=10)	5% (N=1)	36% (N=10)	35% (N=6)
<b>Neither</b>	28% (N=8)	65% (N=11)	21% (N=6)	41% (N=7)
<b>Total Agree</b>	36% (N=10)	30% (N=5)	43% (N=12)	24% (N=4)

Table 2.4: A summary of CanVEC self-reported behaviours and attitudes towards language mixing

<sup>17</sup>There was one participant in this subset, who scored 5 (maximally positive) for all adjective pairs, in both languages. While this might accurately reflect the speaker's opinion, it is also possible that they did not want to make an effort to answer the questions. In either scenario, this data point is an outlier and was therefore removed from the calculation of the mean to avoid skewing the results.

It is also apparent from [Table 2.4](#) that while a large proportion of speakers still agree that languages should be kept separate or that they themselves keep the languages separate, more than half (ranging between 57% and 76%) either do not have a clear opinion or disagree. This suggests speakers' conception of an emergent speech repertoire, i.e. the totality of linguistic varieties at the disposal of, and used appropriately by, a particular speaker (Trudgill, 1974; Platt & Platt, 1975). In other words, the community utilise English and Vietnamese together as their own 'repertoire,' without the need to draw hard and fast boundaries between the languages. It is worth stressing though that this attitude does not tell us anything about the transmission of the heritage language to younger generations. As [Chapter 3](#), §3.3.2 will show, Vietnamese remains the main language of communication within the Canberra Vietnamese community (54%, N = 7,508), with first-generation speakers actively making second-generation speakers use Vietnamese on a regular basis (examples (2) and (3) above).

## 2.5 Chapter summary

In this chapter, I have described both the social and the linguistic background of the community, its speakers and the contact setting in which Vietnamese and English co-exist. Despite a lack of some of the traditional characteristics of a 'speech community,' the Canberra Vietnamese diaspora exhibits other community markers such as a strong social network, shared communal practice, and close personal ties. Linguistically, Vietnamese is the second most popular heritage language spoken at home in Canberra, yet its status as a language is still marginalised, both in comparison to English, and to other 'high-priority' community languages (e.g. Chinese, Arabic, Japanese, Indonesian, German and French) that are believed to bring about better socio-economic benefits.

My direct source of data in this study comprises first and second-generation speakers of Vietnamese in the Canberra community. They represent a range of demographic backgrounds, allowing extralinguistic factors on language variations to be assessed. While the sample speakers' demographic information might seem to be at odds with the general Vietnamese immigrant community overall, their distribution aligns almost perfectly with the Canberra Vietnamese demographic. This shows that the sample is highly representative of the community. In the next chapter, I discuss how these speakers enable us to build a digitalised corpus of the Canberra Vietnamese-English vernacular.

# BUILDING THE CANBERRA VIETNAMESE-ENGLISH CORPUS (CANVEC)

---

## 3.1 Introduction

In this chapter, I build on the previous background to create a new resource that captures the community's vernacular. In particular, I introduce the Canberra Vietnamese-English corpus (CanVEC), an original corpus of natural speech produced by 45 Vietnamese-English bilingual speakers in Canberra, Australia. Here I describe the key components of the corpus, including the data collection process (§3.2) and a new method to semi-automatically annotate the data with language marking, POS-tags and translations (§3.3). Not only does this corpus serve as the basis for the analyses that follow, it also contributes a novel method for semi-automatically processing mixed-language corpora in general. Ultimately, this chapter responds to the first objective of the research, i.e. to create the first Vietnamese-English bilingual corpus that captures the community's vernacular.

## 3.2 Building CanVEC

### 3.2.1 Recording procedures

My principle in building CanVEC was drawn from Labov's emphasis on the vernacular, where 'minimum attention is paid to speech' (Labov, 1984, p.29). The vernacular is particularly suited to the aim of the present study, as it reflects the most natural, systematic form of the language acquired by the speaker 'before any subsequent efforts at (hyper-)correction or style shifting are made' (Poplack, 1993, p.252). To maximise the informal environment which is conducive to

the vernacular, I chose self-recording over sociolinguistic interviews, a popular method used in building other naturalistic corpora (e.g. Tuc, 2003; Nagy, 2011; Torres Cacoullós & Travis, 2015).

Participants in the study were asked to self-record on their mobile phones (a) conversation(s) of a minimum of 30 minutes, with any other bilingual Vietnamese-English speaker. The interlocutor was to be someone the speaker would normally speak with casually, for example a close friend, a colleague or a family member. This led to several speakers asking if their children might participate. Given that young children's speech has long been given separate merit in the literature due to intervening factors of acquisition and developmental stages, I initially rejected the inclusion of children in the recording. Some speakers, however, pointed out that their children fit the description of an 'ideal interlocutor' given their fluency in both languages and their regular interaction with the first speakers. This was taken into account, and upon further consideration of previous work on child language acquisition, a cut-off age of 10 years was subsequently applied. It was rationalised on the basis that, while certain aspects of language have been shown to not be fully acquired until puberty (e.g. see Champaud & Bassano, 1994 for work on discourse markers or Trueswell, Sekerina, Hill & Logrip, 1999 on appropriate use of temporal and discourse connectives), we nonetheless still do not know the exact age for complete maturation of such aspects. The benchmark of 10 years was therefore, although somewhat arbitrary, sufficiently reasonable to ensure that any peculiarities in language patterns could be fairly attributed to community-specific patterns of language interaction and not to 'divergent' child language competence. In fact, given that Pete, the youngest speaker of CanVEC (Table 2.2), is two years above this benchmark, we can reasonably assume that child development as a factor can be ruled out.

Briefing of participants prior to the recording took place in several forms: in person at community events, via emails, text messages and phone calls. Basic information about the study was formally supplied in the form of a bilingual sheet (Appendix C). This sheet introduced the project as studying how Vietnamese bilinguals in Canberra interact, with bilinguals simply defined as those who use Vietnamese and English regularly. No instruction was given to influence whether participants speak both languages in the recording; instead, they were encouraged to converse as they normally would. Two speakers directly asked if Vietnamese or English was preferred, and were told that it was entirely up to them, as long as they talked in the way they normally did. No topics or explicit mention of language mixing was given.

To maximise data authenticity, I was not present during the recordings. Speakers were asked to record themselves using their personal mobile phones. This was methodologically strategic, as a speaker's mobile phone is a familiar item in everyday life and might therefore substantially lessen the intrusive effect that an unfamiliar recording device would have produced. Two participants in their 60s did not own smart phones, however, so were instead given a Zoom H550002

recorder. Most recordings were of high quality, and only one sound file (Tony.Harper.0612) was considered unintelligible and therefore discarded from the corpus. Since speakers had a pre-existing connection with each other, the conversations flowed naturally from the beginning of the recordings, as there was no initial awkwardness. As I show in §3.3.1.3, speakers also discussed highly-sensitive topics which probably would not have been spoken about if they had felt self-conscious.

At the initial stage of the recording process, a number of speakers reported difficulties having a conversation with their partners for 30 minutes continuously. In other studies that utilised sociolinguistic interviews (Torres Cacoullos & Travis, 2015) or pre-arranged recording sessions (Deuchar et al., 2018), this was not an issue. A possible explanation for this difficulty thus could be due to this study's particular emphasis on 'a natural, relaxed chat.' Specifically, since speakers were not put in a conversation with someone unfamiliar (e.g. the researcher in sociolinguistic interviews), or in an artificial environment where the conversation took place (e.g. a recording studio), there was less pressure to 'fill in the gap.' Furthermore, they could also be easily distracted by the other things going on in their familiar environment (mostly their homes), such as getting a drink (Tim.Jess.0708) or answering phones (Tee.Taz.0905) for example. While this set-up is designed to make the data as uncontrived as possible, it might not have been conducive to a continuous dialogue.

Based on feedback from participants, I accordingly adjusted the requirement from having 'a conversation of at least 30 minutes' into 'one or two conversations totalling at least 30 minutes, with no single recording shorter than 15 minutes.' Naturally, participants did not always strictly adhere to instructions, and several conversations returned were still a little shorter than 15 minutes, with the shortest (Tim.Jess.0705) running to 13 minutes and 8 seconds.

I transferred all the recordings onto my computer and a password-protected OneDrive folder provided by the University of Cambridge. Recordings were numbered in the order in which they came in (Table 2.2) and saved as waveform (.wav) sound files compatible with the transcription software ELAN. The procedure generated a corpus of 10 hours and two minutes, and of approximately 90,000 words. As Table 2.2 has already shown, the corpus consists of 23 conversations by 45 Vietnamese-English bilingual speakers (16 of which were cross-generational, seven between Gen 1 speakers, and one between Gen 2 speakers), ranging in age from 12 to 67. Further information on each recording is provided in Appendix D.

### 3.2.2 Questionnaire

After receiving the recordings, I sent speakers a follow-up questionnaire to obtain extra-linguistic information (Appendix E). The questionnaire was available both online and in paper form to

avoid bias towards a particular social group. Speakers let me know via phone or email which version they would prefer to use. While most speakers chose the online version (N=40), some did ask for the paper version (N=5). Those speakers were given the questionnaire either in person or via post with a pre-paid return envelope.

The primary aim of the questionnaire was to gather data on independent variables which would be used to understand variation in the data. The questionnaire was, with slight modification, modelled on one designed by Deuchar et al. (2018) in building their bilingual Welsh-English corpus *Siarad*. The questionnaire was given to speakers both in English and in Vietnamese, and speakers were free to choose to answer in whichever language they were most comfortable with. Out of 45 participants, only eight chose Vietnamese over English. Although this might seem unexpected given that Vietnamese is the preferred language in the speech corpus, it is not inexplicable: English is the majority language, and often the ‘paperwork language’ in speakers’ daily lives.

Since I would not be present at the time the participants filled in the questionnaires, I had conducted two pilot runs to ensure the questions were as well-formulated as possible. Five native English speakers and five native Vietnamese speakers from my informal network in Cambridge participated in this pilot phase.<sup>18</sup> As Adams & Cox (2008) note, the challenge was to ‘strike a delicate balance between collecting as much valid information as needed and keeping questions as short and simple as possible’ (p.18). With this in mind, the final version of the questionnaire consisted of 18 questions, taking into account feedback from the pilots. The questions probed speakers’ demographic information, their self-assessed proficiency of English and Vietnamese, their language attitude as well as the language of their social network. Together with content analysis from the recordings, this allowed appropriate construction of speakers’ sociolinguistic profiles, previously described in [Chapter 2, §2.4](#).

### 3.3 Annotating CanVEC

Annotating and organising the collection of a new speech corpus is known to be an intense endeavour in terms of both time and labour, and requires careful consideration of numerous practical decisions and theoretical assumptions (Caines, Bentz, Graham, Polzehl & Buttery, 2016; Bullock, Serigos, Toribio & Wendorf, 2018b; Torres Cacoullous & Travis, 2018; Deuchar et al., 2018). In this section, I thus describe three key components of the annotation process for CanVEC: transcription conventions (§3.3.1), data processing (§3.3.2) and evaluation procedure (§3.3.3). A full summary of the conventions used can be found in [Appendix F](#).

---

<sup>18</sup>Note that due to funding restrictions, none of the speakers in this pilot phase was paid.

### 3.3.1 Transcription method

The first step in the annotation process is transcription. The choice of transcription conventions must be suitable for the purpose of the research, requiring decisions regarding units of analysis, levels of phonetic detail, levels of non-linguistic marking and so forth. A study taking a conversational analysis approach, for example, would require detailed marking of all interactional cues such as pauses, false starts, laughter or overlaps. In the context of CanVEC, such details are not required. Instead, for a purpose-built corpus for the study of languages in contact, the identification of language membership (i.e. language marking) of each token and clause is a much more crucial part of the transcription.

#### 3.3.1.1 Sound to text

All transcriptions of CanVEC are time-aligned, which means each specific stretch of speech is linked up to its corresponding texts, with a specific time stamp. Although time-aligned transcriptions are more time-consuming than text-only transcriptions, the direct link between the text and the recording offers some clear advantages: easy access to the original audio associated with a stretch of written record, customised tiers for interlinear glossing and tagging, and easy conversion of data into displayable formats for presentation (Thieberger & Berez, 2012).

Data in CanVEC was transcribed using ELAN, a transcription software developed by the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands (Sloetjes & Wittenburg, 2008).<sup>19</sup> Key features of ELAN include, but are not limited to, the ability to segment utterances and separate them into linked tiers. The main tiers allow transcription of features in natural speech such as pauses, repetitions, interruptions, and overlaps between speakers.<sup>20</sup> The software also enables sophisticated searches, concordance, and statistics regarding frequency of occurrence. Transcription filenames were given in the format **Speaker 1's Pseudonym. Speaker 2's Pseudonym. Date Recorded**. For example, **Tim.Jess.0704** indicates a transcribed conversation between Tim and Jess, on July 4th.

To facilitate automated language marking at a later stage, I used standard English orthography for phonetically realised English words and Vietnamese orthography for phonetically realised Vietnamese words. Example (4) below provides an illustration:

- (4) Tanner<sub>1</sub>: price *chỗ* *đó* *nó* good *ha*  
                   place DEM 3SG           DM  
                   'As for price at that place, it is good.'

(Tanner.Nina.0609. 01:10.9–01:13.7)

<sup>19</sup><https://tla.mpi.nl/tools/tla-tools/elan/>

<sup>20</sup>While it is not pertinent to the present study to annotate such details, these options allow future studies to explore the data further.

As example (4) shows, all Vietnamese standard orthography and diacritics are respected for words pronounced in Vietnamese, and so is English standard orthography for words pronounced in English. Since not all Vietnamese words have diacritics, naturally there are overlaps of orthography between two languages such as *ha* in (4) above. These cases will be further distinguished in §3.3.2.2, where I discuss language-marking for ambiguous items.

It is important to note that although all speakers are part of the Vietnamese community living in Canberra, they originally come from different parts of Vietnam. This diversity has created a corpus representing various regional dialects, consisting of Northern, Central, and Southern varieties. To maximise data consistency, I opted for standard Vietnamese orthography for all phonetic variants (Poplack, 1993, p.265), a practice previously adopted in Nguyen (2016). For example, in Vietnamese, the onset /v/ has two variants including [j] and [v]. It was common for participants from Central and Southern Vietnam to pronounce the alveolar fricative [j] instead of the labial fricative [v] in words such as *bởi vì* ‘because’ or *sao vậy* ‘why’ in informal speech. All of these variants were transcribed as the standard <v>. Note that this has no adverse implications for the study as the standardisation only concerns phonetics; all lexical and syntactic variations are kept as originally produced by the speakers. On a larger scale, orthographic consistencies also bring about crucial benefits in enhancing searchability, enabling automated treatment of the corpus and facilitating computer-assisted analysis.

It is also commonly agreed in corpus linguistics that accurate transcriptions require multiple rounds of revisions (Nagy & Sharma, 2013; Torres Cacoullós & Travis, 2018; Deuchar et al., 2018). High-quality published corpora all involved extensive labour from multiple transcribers over the course of several years. Some examples include the Ottawa-Hull French corpus (Poplack, 1989), the Multilingual London English (MLE) corpus (Cheshire, Kerswill, Fox & Torgersen, 2011), the Heritage Language Variation and Change (HLVC) corpus (Nagy, 2011), the New Mexico Spanish English Bilingual (NMSEB) corpus (Torres Cacoullós & Travis, 2018), or the Siarad corpus (Deuchar et al., 2018). Although it was not feasible to uphold the same standard due to the time and resource constraints of this project, I conducted transcriptions with the same principles in mind. Specifically, I transcribed all of the recordings twice, with at least a week in between the first and the second pass. The gap of time between the two rounds was to ensure that transcriptions were done appropriately under different settings, at different times, with fresh eyes and mind. The ultimate aim was to minimise human errors, thereby creating a dataset as accurate as possible.

Roughly 10% of the data (i.e. a random chunk of 10 minutes each in six different conversations) was also additionally annotated by a second transcriber to further enhance transcription reliability. Although the primary researcher’s direct and constant engagement with the corpus is

crucial in the data analysis process, transcriber effects are unavoidable. As Jung & Himmelmann (2011, pp.208–209) pointed out, transcribers who are community members, either consciously or unconsciously, often resist transcribing verbatim certain elements of the recording due to taboo, disbelief, or natural concern for the message rather than for the form of the utterances. Chunks of audio material could also be easily missed due to the human facility of attending to salient constituents of the message and tuning out those perceived to be irrelevant (Nagy & Sharma, 2013, p.253). To minimise this effect, I engaged the help of a linguistics student, who is a native English speaker and fluent in Vietnamese to perform a reliability check. It was important that the assistant’s primary competence was in English rather than in Vietnamese, as the second transcriber was more likely to catch words in their native language than the first transcriber (whose native language is Vietnamese) might have missed or misheard (see Torres Cacoullos & Travis, 2018). The benefits of this method became clear in the process of transcribing CanVEC, as the assistant was able to pick up one or two English words that I previously had not been able to decipher (e.g. ‘professional gymmer’ in ‘Therese.Luna.0703’).

As previously indicated, other than standardising phonetic variants, no other effort was made to correct the form of speakers’ speech in any way. An overall rule for both transcribers was to prioritise accuracy as there would be no point collecting the vernacular without appropriately reproducing its elements in the transcripts. Lexical choice, deletion, disfluency, and syntax were thus all faithfully preserved. Deletion, however, has been noted to pose a challenge in transcription. Poplack (In press, p.8), for example, notes that given ‘the daunting number of disparate forms that would have to be coded as null, coupled with the difficulty of finding a unique representation for each (one capable of distinguishing a null subject from a null complementiser or a null inflection, for instance), eventually led to a point of diminishing returns.’ Deletion was therefore not marked during the transcription, but later manually coded instead (Chapter 5, §5.5).

Despite the overall high quality of the recordings, unclear speech features appear occasionally in the corpus. In accordance with the method by Du Bois, Schuetze-Coburn, Cumming & Paolino (1993), each unclear syllable was marked with an <X>’. There were also instances where, even though the syllables were unclear, I had a good idea of which language the relevant segment had been expressed in. These clauses were then treated according to the transcribers’ ‘best guess’ (Du Bois et al., 1993, p.75). That ‘best guess’ was incorporated into the transcription using angle brackets, and was marked as <V> if it was considered more likely to be Vietnamese, or <E> if it was considered more likely to be English. Examples (5) and (6) demonstrate:

- (5) a. Naomi<sub>2</sub>: well no but it was quite a challenge for me.  
 b. today when we were doing mental <E>,  
 c. that was just <E>.

(Lina.Naomi.0623, 09:23.5–09:33.9)

- (6) a. Tom<sub>1</sub>: *mà hình-sự nó cho* prosecutor *điều-tra là* <V> *ấy,*  
 but criminology 3SG let investigate COMP DM  
 ‘(If) the police send prosecutors to investigate, then <V>.’
- b. <X> *chơi kiểu đấy đâu*  
 play type DEM NEG.DM  
 ‘X won’t behave like that.’

(Tom.Henry.0725, 42:22.7–01:13.7)

Following Deuchar et al. (2018), I used Turnitin<sup>21</sup>, a commercial plagiarism detection service, to measure the overlap between my transcription and that of the second transcriber. The software compared the two versions of the transcriptions and calculated the overall similarity (%) between the two texts. Documents were then returned with highlighted annotations, showing where similarities and differences occur. As Turnitin reported, the matching rate was exceptionally high, reaching 95% overall. Most of the differences identified were typos, spelling errors or incorrect display of the Vietnamese diacritics. These were fixed accordingly for the final analysis. The use of Turnitin hence also aided the identification and correction of transcription errors in the corpus, which ultimately helped improve its overall reliability.

### 3.3.1.2 Segmentation: Unit of analysis

After audio data has been transcribed into text, a crucial step in transcription is segmentation, i.e. the process of splitting the stretches of utterance into consistent boundaries such as turns, clauses, or intonation units. As speech does not contain any explicit boundary markers, e.g. punctuation, this requires careful consideration; there is a trade-off between the granularity and the versatility of the transcription. For the purpose of data processing in particular, word-level segmentation may be better for speech recognition systems, but POS-tagging and parsing work best at the sentence level. In CanVEC, this is even more challenging since spoken Vietnamese deviates significantly from the standard written form, and spoken language in general naturally contains fragments, disfluency and false starts.

Given that speech has been shown to be non-sentence-based, co-constructed and highly-interactive (Carter & McCarthy, 2017), the first round of CanVEC segmentation involved roughly dividing speakers’ speech into turns, and then by Intonation Unit (IU), which is defined as ‘a sequence of words combined under a single, coherent intonational contour’ (Chafe, 1987, p.22). As Chafe (1994) explains, speakers are often unable to process large amounts of information in active consciousness at any given time, and thus tend to break ideas into ‘functionally relevant segments of speech’ (p. 57). Each of these segments then activates a new piece of information in terms of attention focus and carries one single new idea. From this functional perspective,

<sup>21</sup><http://www.turnitin.com>

Chafe argues for intonation contour as a robust indication of a fast and firm boundary, marking a locus for the cognitive processing of the preceding information.

It is important to note that although IU is conceptually close to other prosodic units, such as ‘tone group’ or ‘information unit’ (Halliday, 1967), ‘tone unit’ (Brazil, 1985) or ‘idea unit’ (Chafe, 1980), it is formally distinguishable from all of these. The key difference that sets the IU apart is that it is identified based on the boundaries of the unit, while other concepts use the internal structure of the unit. The criterion for identifying an IU is a ‘convergence of prosodic cues’ (Travis, 2005, p.22), of which a ‘coherent prosodic contour’ is the foremost qualifying criterion. The cohesion of a prosodic contour could be characterised by change in pitch reset, changes in word duration (perceived as lengthened IU end or rushed IU initial), change in intensity (recognised as loudness), pauses, or changes in voice quality (often perceived as creak) (Du Bois et al., 1993; Chafe, 1994). Even though not all of these prosodic features must be present, it is crucial that these cues are used together to identify IUs in the dataset. Reliance on only one cue could be misleading; for instance, while pauses often delimit IU boundaries, it is not unusual for them to be found within the IU (Du Bois et al., 1993; Chafe, 1994; Travis, 2005; Shenk, 2006).

In the context of CanVEC, however, it is important to acknowledge that it was not always straightforward to gather several cues for an IU boundary. As Vietnamese is a tonal language, pitch reset can be obscured by tonal information and does not always signal the beginning of a new IU (Li, 2014). A lengthening IU could simply be a manifestation of disfluency and hesitation rather than a marker of a complete IU. These difficulties, however, are not new, and have been recognised in previous studies of IU in tonal languages (see Tao, 1996; Li, 2014 for Mandarin Chinese and Nguyen, 2018 for Vietnamese). It has been established that the movement of a prosodic unit, rather than the contour shape, is a better identifier of intonation patterns when lexical tones are in play.

In a previous study on Vietnamese (Nguyen, 2018), I showed how the falling intonation contour can appear fairly frequently, and pitch reset acts as a reliable cue for IU identification in Vietnamese. This observation generally repeats in CanVEC. [Figure 3.1](#) exemplifies a typical case.

Here, we can see that an obvious contour is clearly present in monolingual Vietnamese. In some instances, such as *hắn đi ra* ‘(when) he goes out’, we also see that the pitch of the IU begins higher and reduces over time to an eventual drop. It therefore can be said with reasonable confidence that despite the interplay of tones and local pitch at a lexical level, pitch reset at the level of speech stretch generally exists. This primary evidence is considered along with secondary cues wherever possible, such as pauses and changes in voice quality. Modelling on Du Bois et al. (1993), basic delimiters such as final prosodic contour (i.e. when the speaker intends to stop) are marked either with a full stop (falling pitch) or a question mark (rising pitch), and non-final

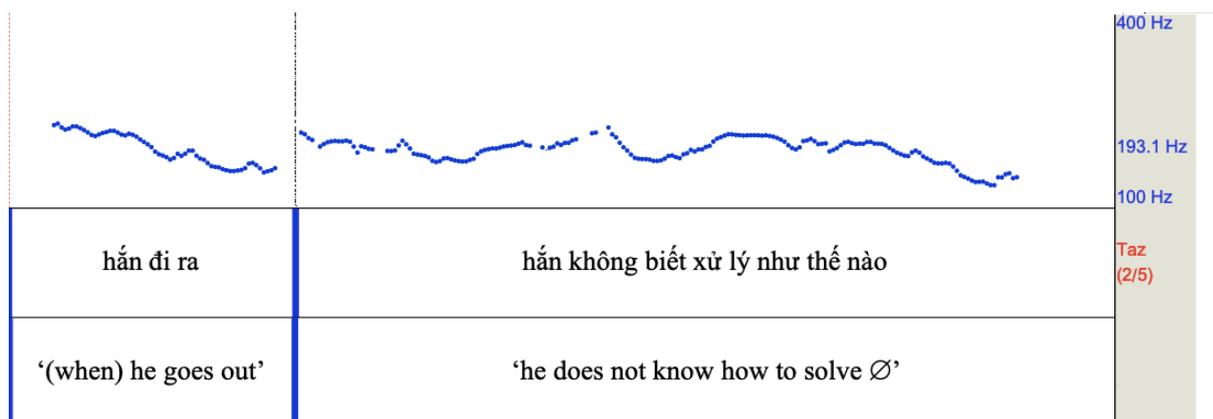


Figure 3.1: A Praat demonstration of Vietnamese pitch reset at IU boundaries (Tee.Taz.0808, 00:28.4 00:30.5)

contour (i.e. when the speaker intends to continue) is marked with a comma (a slight rise in pitch). This is marked at the end of each IU, which is given on a separate line of the transcript. Example (7) illustrates this system.<sup>22</sup>

- (7) a. Tyler<sub>2</sub>: we never have conversations,  
 b. we just stare at each other.  
 c. Billy<sub>1</sub>: we never did?

(Billy.Tyler.Ellie.0807, 17:07.2–17:16.0)

It should be made clear is that though it is a prosodic unit, the IU is relevant for the linguistic construction of an utterance. The correlation between prosody and syntax has long been noted (e.g. Chafe, 1994; Ford & Thompson, 1996), and a robust correspondence between an IU and a grammatical unit has been consistently shown (Ford & Thompson, 1996; Shenk, 2006; Torres Cacoullos & Travis, 2015). In Ford and Thompson's (1996, p.155) study, for example, almost 100% of the prosodically complete IUs (N=433) were found to be 'syntactic completions' (N=428). Torres Cacoullos and Travis (2015) also found that 95% (N=664/696) of pre- and post-verbally expressed first-person singular subject pronouns in their dataset occurred in the same IU as the verb. Similarly, Shenk (2006) observed that there are no instances of an object occurring in a different IU to the main verb.

In CanVEC, when a subject is present, the subject and its immediately following verb are almost never split across two IUs (barring coordination constructions, as in lines (f.)–(g.) in example (8) below). The following examples (8) and (9) illustrate how each IU corresponds with a syntactic unit (in most cases, a clause):

<sup>22</sup>Truncation and other fine prosodic details were not marked, primarily because they are not particularly pertinent to the analysis and also due to lack of time available.

- (8) a. Reece<sub>1</sub>: that is what interesting,  
 b. and Vietnamese woman no they do not have to wait for the cubicle.  
 c. no need to go,  
 d. Taylor<sub>2</sub>: so they don't have to wait to go into the cubicle?  
 e. Reece<sub>1</sub>: they just stand in there,  
 f. get the water,  
 g. and shower themselves.

(Reece.Taylor.0906, 53:35.0–53:45.9)

- (9) a. Taz<sub>1</sub>: *em nghĩ là,*  
 1SG.kin think COMP  
 'I think that,'  
 b. *mình chưa có do enough cho hắn.*  
 1PL NEG have for 3SG  
 'We haven't done enough for him.'  
 c. *nó tới cái stage,*  
 3SG reach CLS  
 'He reached a stage,'  
 d. *mình run out of time rồi.*  
 1PL PERF  
 'We have run out of time.'

(Tee.Taz.0808, 08:23.6–08:37.0)

As these examples clearly demonstrate, there is a strong prosodic tendency to keep clausal elements together in the same IU.<sup>23</sup>

Prosody, furthermore, plays a crucial role in enabling the researcher to fully understand what is being said (Torres Cacoullous & Travis, 2015; Torres Cacoullous & Travis, 2018; Nguyen, 2018). Consider examples (10) and (11) below. As the transcription stands in (10), the 2SG kin term *anh* could be interpreted as either a second-person singular subject for the VP *không biết* (don't know), or a vocative following the whole clause *nó không biết* (s/he doesn't know). It is only when the prosodic boundaries are marked as in (11), that the clauses can be delimited: *anh* is a vocative instead of an argument, and the verb 'know' was used as an intransitive verb.

- (10) Ellie<sub>1</sub>: *mà nó không biết anh*  
 but 3SG NEG know 2SG.kin  
 'But he doesn't know you'

(Billy.Tyler.Ellie.0807, 02:23.4–02:24.3)

<sup>23</sup>Those who adopt the generativist Y-model (also discussed in Chapter 6) would prefer an alternative interpretation: clausal structures determine prosodic boundaries. In other words, prosody only reflects an interpretation of syntactic structure. As Theresa Biberauer (p.c.) points out, however, the Y-model is not a psycholinguistically oriented model; it is about how different aspects of language structure connect up, and as such we need to consider, as a separate question, how the Y-model can be integrated into a plausible, temporally oriented and not just a 'unidirectional' psycholinguistic model, in which the phonological and logical form are straightforwardly computed post-syntactically. Whichever of these views is preferred, nonetheless, the important point to stress here is that there is a strong correlation between syntax and prosody in speech. As prosodic boundaries are more salient in natural data, they were selected as the cues that would serve as the basis for the delimitation of spoken (syntactic) units.

- (11) a. Ellie<sub>1</sub>: *mà nó không biết,*  
           but 3SG NEG know  
           ‘But he doesn’t know,’  
       b. *anh.*  
           2SG.kin  
           ‘You.’

This is of particular relevance for the present study, as prosody provides helpful cues to accurately determine the syntactic role of each element. In the above examples, the prosodic break specifically enables us to understand the syntactic role of *anh* as precisely intended by the speakers.

It is, however, worth recognising that while an IU often corresponds to a point of ‘syntactic completion,’ it may also consist of non-clauses (line c. in 12) or multiple clauses (line d. in 12).

- (12) a. Jess<sub>2</sub>:           on my <Explore page> on Instagram I just saw these boys,  
       b. Chloe<sub>1</sub>:       and you fell in love.  
       c. Jess<sub>2</sub>:       yeah love at first sight.  
       d.               before that I thought like Kpop was really bad.

(Tim.Jess.Chloe.0705, 08:32.1–08:53.0)

Since one of the aims of the dissertation is to examine the Matrix Language Turnover Hypothesis (Chapter 4) which takes clause as a unit of analysis, these cases require a second round of segmentation. Specifically, non-clausal IUs such as line (c.) in (12) are marked for exclusion, and split multi-clausal IUs such as line (d.) in (12) into two separate clauses. In other words, line (d.) in example (12) will become two units of analysis, represented as line (d.) and (e.) in example (13). Relative clauses are not separated, but grouped together with the main clause whose components they modify (Hurewitz, 1998). Each clause is then represented by a separate line on the transcript.

- (13) d. Jess<sub>2</sub>:       before that I thought  
       e.               like Kpop was really bad.

While this extra step might seem to further complicate the consistency of a ‘unit of analysis,’ this laboriously fine-tuned delimitation is justified. It is worth emphasising that since most IUs correspond to a clause already, the unit of analysis in CanVEC remains fairly consistent, which can be dubbed **IU-approximates**, in this case, a clause. It is also worth noting that clauses further segmented in the second round, such as line (d.) in example (13), do not have IU-boundary markers (i.e. comma for continuing intonation, period for a full stop, or question mark for a rising continuation, as previously described). This is because there is no prosodic break between the clause and its following one, in this case, line (e.).

### 3.3.1.3 Ethical considerations

One of the main commitments of this study is to make CanVEC accessible to future research. This is motivated both by Labov's (1982) 'Principle of the debt incurred'<sup>24</sup>, and by the overall lack of comparative data on under-described language pairs in communities outside a European or Western context. It is, however, also linguists' responsibility to protect the community and ensure that their data is appropriately handled (Travis & Torres Cacoullos, 2013; Torres Cacoullos & Travis, 2018). Several ethical considerations are thus worth addressing in relation to data management.

First of all, the speaker in charge of the recording was advised that the other speaker must be aware that the recording was taking place, and that signed consent must be obtained from both speakers before the session (Appendix C).<sup>25</sup> The emphasis on natural settings, however, meant that more speakers might sometimes randomly join the conversation. This occurred in five of the recordings, and all extra speakers later gave consent to have their data transcribed and analysed as part of the corpus. Seven speakers were under 18, and were asked for written consent both from themselves and from their guardians. I encouraged speakers to listen to their recording afterwards and decide whether there was any portion of their conversation containing private, sensitive information that they would like to delete. Once they returned the recording to me, I initially proceeded on the basis that speakers had agreed for all parts of it to be used for research purposes. However, I soon realised during the transcription stage that several parts of the corpus touched on private topics such as speakers' gambling history, gossip or hypothetical scams. While this is a valuable indication of the naturalness of the data collected, it posed the question of whether all speakers had fully read and understood the terms and conditions that they agreed to. In her work on Palauan English for example, Matsumoto demonstrates this point by quoting one of her participants:

When I went to University X in the US and found out how my relatives were quoted in theses, I was really in shock. You know, there're things I swear by God they would never say openly if they'd known their words would be published with their own names. You know, Americans would've thought that we'd never read their theses.

(Unpublished manuscript, cited in Cheshire & Fox, 2016, p.295)

<sup>24</sup>Labov's 'Principle of the debt incurred' states that 'An investigator who has obtained linguistic data from members of a speech community has an obligation to make knowledge of that data available to the community, when it has need of it' (1982, p.173). It specifies that, while linguists must be fiercely committed to the privacy of their sources, the knowledge that springs from linguistic analysis is in principle the general property of the speech community, and it is in nobody's interest that such property remains buried in the linguist's field notes or unpublished papers.

<sup>25</sup>This also applies to young speakers under the age of 18. In the event that the 'primary speaker' was not their caregiver, consent was additionally sought from their caregivers.

Setting aside the obvious problem of lacking anonymity in this scenario, the participant's comment highlights the needs for researchers to be aware of locally established norms and taboos in the community. While it is not always possible for researchers coming from outside to do so, such a requirement is made possible in the present study due to my status as a community member. With the advantage of knowing and living 'the norms,' I was able to informatively assess the data and take into account appropriate ethical considerations. For example, as Vietnamese culture is highly collective (Parks & Vu, 1994; Carruthers, 2008a; see also [Chapter 2, §2.4.3](#)), concepts such as privacy or anonymity are relatively far removed. First-generation speakers in particular do not fully understand what constitutes 'personal information,' let alone the consequences of their personal information being public. On the other hand, second-generation speakers grew up (or are growing up) in Australia and are less familiar with Vietnamese taboos and cultural etiquette. As a result, they occasionally made sexually or politically sensitive references, which are controversial in certain contexts. These parts will be removed from the open-access corpus. The decision was made with the community's best interests at heart, and is a compromise between the competing needs of releasing the data on the one hand, and protecting the minority communities from reinforced negative stereotypes on the other (Torres Cacoullos & Travis, 2018). Crucially, the protocol in this study is fully GDPR-compliant<sup>26</sup> and furthermore adheres to the strict guidelines of the National Statement on Ethical Conduct in Human Research in Australia, i.e. where the community was based and data was collected.<sup>27</sup>

In the preparation of the transcripts, the identity of all speakers was anonymised. This turned out to be a rather labour-intensive task. Specifically, as Vietnamese speakers regularly use names in place of personal pronouns for self- and interlocutor-reference (Nguyen, 2018), a large number of personal names are present in the corpus. All speakers were thus given a pseudonym, and all third-person individual names mentioned in the transcripts were replaced with a generic '[A:person name]' as demonstrated in example (14). Occasional references to public figures with unfavourable remarks were also treated with caution, in that the name of the person and any associated defining characteristics were anonymised and removed respectively. Any remaining comments are kept if they have then become sufficiently ambiguous in terms of whom they refer

<sup>26</sup>The EU General Data Protection Regulation (GDPR) and the UK Data Protection Act came into effect in 2018. It governs the processing (acquiring, holding, using, etc.) of personal data in the UK. The new law demands that data processing is lawful, fair and transparent. There are 6 **lawful bases**, upon at least **one** of which research must operate. The processing of CanVEC personal data in this dissertation has been verified by the Government interactive tool as meeting two of these lawful bases: Consent and Public interest. For further information about GDPR, see <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/>.

<sup>27</sup>The full National Statement Guidelines can be found at <https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018>.



nations of somewhat well-resource language pairs such as English-Spanish (Solorio & Liu, 2008; Solorio, Blair, Maharjan, Bethard, Diab, Ghoneim, Hawwari, AlGhamdi, Hirschberg, Chang & Fung, 2014; Bullock, Serigos, Toribio & Wendorf, 2018a), English-Hindi (Diab & Kamboj, 2011) or English-Mandarin (Lyu, Tan, Chng & Li, 2015), research involving low-resource, or less-described languages such as Vietnamese is still largely neglected. This means very few off-the-shelf resources are available. For CanVEC particularly, apart from the noises commonly characteristic of speech, the presence of two languages (with a certain degree of orthographic overlap) in the same discourse poses a main hurdle. To pioneer the much-needed work for processing this specific language-pair, and to make CanVEC maximally exploitable for future research, I collaborated with a computer scientist at the Cambridge Computer Lab<sup>29</sup> to devise an algorithm to tackle some parts of this unsolved task. In what follows, I will describe the key component of this process, much of which has already been published in Nguyen & Bryant (2020).

### 3.3.2.1 Automatic language marking and Part of Speech (POS) tagging

Given the contact setting of this study, language marking and POS-tagging are the foremost levels of annotation required for our corpus. First, language marking is key to assigning language membership of individual tokens; and second, it helps determine whether language mixing has occurred in any given clause. POS-tagging, on the other hand, enables us to efficiently identify consistent clause units.

As a result, we developed a Python script that performs the following:

- (i) tokenising Vietnamese items into words;
- (ii) tagging each token in the corpus for language membership;
- (iii) tagging each clause as monolingual or mixed; and
- (iv) tagging each token for parts of speech.

Before elaborating on the workflow, it is worth specifying that task (i)—tokenising Vietnamese items—is particularly important. This is because each graphic unit in Vietnamese corresponds to a syllable (*tiếng*), which may or may not be a complete word. While Vietnamese has been described elsewhere as monosyllabic (Emaneau, 1951; Brunelle & Le, 2014), this label is misleadingly simplistic.<sup>30</sup> In fact, although a minimal lexical item in Vietnamese may consist of one syllable, most words in fact consist of two or more (Nguyen, 1997). These syllables as parts

<sup>29</sup>I thank Christopher Bryant for his assistance with coding and finding relevant libraries for this project.

<sup>30</sup>Julio Song points out that this is also connected to a bigger debate on the various definitions of a word in isolating languages in general. See Packard (2000, chapter 2) for a comprehensive discussion on this topic in the context of Chinese.

of the same word are also separated by whitespace in the writing convention. In order for the POS-tagging to work efficiently then, it is first necessary for these stand-alone graphic units to be tokenised into lexical items.

With clauses already being segmented, we then devised a workflow to achieve the aforementioned objectives. Specifically, we:

1. Removed inconsistently transcribed punctuation and other artefacts from the clause;
2. Split the clause into text units based on whitespace; although whitespace marks word boundaries in English, it marks syllable boundaries in Vietnamese.
3. Tested each word/syllable for language membership using a Vietnamese syllable list and an English word list.
4. Sent the largest contiguous sequence of Vietnamese or English text to the relevant tokeniser and POS-tagger.
5. Redefined the word level language tag in terms of tokens rather than lexical items.
6. Assigned a clause-level language to the properly tokenised clause.
7. Translated Vietnamese to English in the monolingual Vietnamese and mixed clauses.

This process is also illustrated in [Table 3.1](#). When testing for language membership, we compared each whitespace-separated unit against some large lists of valid Vietnamese syllables<sup>31</sup> and English words<sup>32</sup>. We next sent the largest contiguous sequence of same-language units to a Vietnamese or English POS-tagger as appropriate. Note that language-neutral tokens (see §3.3.2.2) were ignored when defining sequence boundaries. Specifically, we used *Underthesea*<sup>33</sup> v1.1.6 to tokenise and POS-tag Vietnamese sequences, and *spaCy*<sup>34</sup> v1.9.0 to tokenise and POS-tag English sequences. These resources were chosen mainly for their versatility and high performance.<sup>35</sup>

After tokenisation, we were also able to update the language tags in terms of tokens rather than units. This could not be done sooner because we previously did not know which monolingual tokeniser a clause or sequence should be processed by. The clause-level language tags were then defined based on the token-level language tags as follows:

---

<sup>31</sup><http://www.hieuthi.com/blog/2017/03/21/all-vietnamese-syllables.html>

<sup>32</sup><https://sourceforge.net/projects/wordlist/files/speller/2017.08.24/>

<sup>33</sup><https://github.com/undertheseanlp/underthesea>

<sup>34</sup><https://spacy.io/>

<sup>35</sup>Since each POS-tagger uses a different tagset, a tag map was defined to convert all POS tags to the Universal Tagset (Petrov, Das & McDonald, 2012). Although *spaCy* includes a function to do this automatically, *Underthesea* does not, so we instead defined our own mapping function ([Appendix G](#)). This mapping ensured POS-tag consistency across the whole corpus.

Step	Description	Example										
1	Data cleaning	I don't không có really hiểu cái point of it										
2	Split on whitespace	I	don't	không	có	really	hiểu	cái	point	of	it	
3	Test language membership using word/syllable list	I	don't	không	có	really	hiểu	cái	point	of	it	
		@eng	@eng	@vie	@vie	@eng	@vie	@vie	@eng	@eng	@eng	
4	Tokenise and POS-tag same-language sequences	I	do	n't	không	có	really	hiểu	cái	point	of	it
		PRON	VERB	ADV	ADV	VERB	ADV	VERB	CLS	NOUN	PREP	PRON
5	Redefine language tags in terms of tokens	I	do	n't	không	có	really	hiểu	cái	point	of	it
		@eng	@eng	@eng	@vie	@vie	@eng	@vie	@vie	@eng	@eng	@eng
6	Assign clause-level language	@mix										
7	Translate Vietnamese and Mixed clauses	I don't really understand the point of it										

Table 3.1: A demonstration of step-by-step automatic annotation using an example clause

1. Language-neutral tokens (@non) are excluded from the analysis.
2. If all remaining tokens are @vie, the clause is monolingual Vietnamese.
3. If all remaining tokens are @eng, the clause is monolingual English.
4. If there is a mix of tokens from both languages, the clause is mixed (@mix).
5. Otherwise the clause consists entirely of language-neutral words (@non).

Additionally, recall from §3.3.1.1 that unintelligible tokens were marked as <X> during transcription, and those that were more likely to be English were <E>, and those more likely to be Vietnamese were <V>. Given that one syllable may correspond to one lexical item in either language, any monolingual clause with an <X> was assigned <X> overall to avoid doubts around language of the clause, or in other words, whether language combination occurred in that utterance or not. This is illustrated in example (15).

Speaker	Clause	Clause Language
(15)	a. Ellie: <i>điện-thoại</i> <i>rằng</i> <i>hắn</i> <X>.	<X>
	telephone why 3SG	
	why is the phone <X>?	
b.	X,	<X>

(Billy.Tyler.Ellie.0807, 08:23.7–08:28.7)

In case a whole utterance is unintelligible as in line (b.) of example (15) above, it was automatically language-tagged as <X> by the script (N=178). All <X> clauses were subsequently excluded from the analysis.

### 3.3.2.2 Manual verification: Language-neutral items, non-linguistic items, and established borrowing

When testing language membership of each token against valid syllable lists (step 3 in Table 3.1), units that appeared in both or neither lists were held aside to be resolved manually. A large number of these ambiguous units were in fact proper nouns, interjections and fillers, such as ‘uhm’ and ‘okay,’ which are not exclusive to any language, and were therefore marked as language-neutral (@non) (Riehl, 2005). These tokens were subsequently ignored in the following steps, including tokenisation, POS-tagging, and assigning clause-level language. Example (16) provides a case in which a proper noun appears in an otherwise English clause. This was not considered to affect the language of the clause, which was then coded as English monolingual:

	Speaker	Clause	Clause Language
(16)	a. Reece <sub>1</sub> :	we all left Saigon alright?	English
	b.	we call Saigon,	English
	c.	we do not call Ho Chi Minh city okay?	English

(Reece.Taylor.0906, 00:51.9–00:57.9)

The remaining ambiguous units, such as ‘so,’ which means ‘to compare’ in Vietnamese, but is a conjunction in English, were otherwise fairly rare. In any case, these units were still verified against the sound file to determine language membership. Since words often have quite distinct vowel quality in Vietnamese as compared to English, their phonetics can help disambiguate orthographically confusing cases. For example, the token ‘so’ would be tagged as English if it was phonetically realised as /səʊ/, and Vietnamese if it was /sə/.

Having dealt with these superficially ambiguous cases, we addressed a more controversial, yet important aspect of the corpus: the distinction between borrowing versus code-switching. Setting aside the controversy of whether or not borrowing and code-switching are separate processes, researchers have generally agreed that long-term borrowing is well-integrated into the community and thus forms part of monolingual speech.<sup>36</sup> Language marking in a bilingual setting where two varieties are in contact therefore requires extra caution. Other than language-neutral tokens automatically singled out against the online lists, we essentially need to account for established borrowing, which is ascribed to the recipient language. Traditionally, this is often

<sup>36</sup>Borrowing versus code-switching is the subject of a longstanding debate in code-switching research. To summarise, some researchers have proposed that code-switching and borrowing are essentially similar phenomena lying along the same continuum of language contact, evolving from code-switches to established borrowings (Gardner-Chloros, 1991; Myers-Scotton, 1993; Winford, 2003; Treffers-Daller, 2005; Gardner-Chloros, 2009; Winford, 2009), while others believe that they are distinct processes and efforts should be made to distinguish them (Poplack, 1980; Aaron, 2015; Nguyen, 2016, 2018; Torres Cacoullous & Travis, 2018). Whether or not they need to be differentiated, and if so how, remains largely controversial. The criteria proposed and ways to apply them in treating single other-language items also vary, ranging from frequency, diffusion, dictionary attestation to integration and many more.

dealt with using dictionary attestation: if a single word in language A can be found in a dictionary of language B, it is considered a borrowing and hence given language B membership alongside language A. This method is not flawless, however, first because dictionaries often lag behind contemporary usage, and second because the criteria for warranting a word an entry in a dictionary are not always explicitly explained (and therefore poorly understood). Furthermore, for an established bilingual community that is far removed from the homeland, the only appropriate dictionary to consult would be a regional dictionary, compiled for and by community members. Unfortunately such a resource is non-existent, leaving the closest references the Oxford Australian English Dictionary which is rather broad, and the normative Vietnamese dictionary which is compiled for and by people living in Vietnam.

For these reasons, I identify established borrowing using a frequency and diffusion measure instead. Following Poplack and associates (1988, p.52), ‘frequency’ refers to the number of tokens occurring in the corpus, while ‘diffusion’ refers to the number of different speakers using that item. Frequency and diffusion should be treated as two separate criteria, and only when combined can we establish a solid indication of established borrowing (or lack thereof). In their study of 120 speakers in a French-Canadian bilingual community, for example, Poplack, Sankoff & Miller (1988) established four levels of frequency and diffusion: nonce, idiosyncratic, recurrent and widespread. A nonce item is operationalised as a single other-language word that is used only one time in a given corpus, while items used more than once by just one speaker are idiosyncratic. Lexical items that occur more than 10 times are recurrent, and those that are used by more than 10 speakers are widespread. For the purpose of identifying established borrowing here, I only consider items that are both recurrent and widespread. However, relative to the sample size, I redefine ‘widespread’ as items that are used by more than five speakers.<sup>37</sup>

Accordingly, all mixed clauses with a single other-language item were extracted (N=1,904).<sup>38</sup> Their frequency and diffusion level was then quantified in Python, and only those that were both recurrent (>10 times in the corpus) and widespread (>5 speakers) were marked for established borrowing status. As a result, 821 different types were reported, 13 of which were ‘frequent,’ and 14 were ‘diffuse.’ Table 3.2 lists all tokens that are at least frequent or diffuse, with the left half singling out those that meet both criteria. These items were then language-tagged as their surrounding language, which in turn renders their clauses monolingual clauses.

<sup>37</sup> Given the size of the corpus, one might suggest a hybrid approach which makes use of both a dictionary and word frequency/diffusion. The main prerequisite for such an approach, however, remains a regional dictionary compiled for and by community members (i.e. contact speakers). Unfortunately, as I briefly mentioned above, such a resource does not exist and so this approach is not currently feasible.

<sup>38</sup> Note that Vietnamese kin terms used as self- and interlocutor-references were excluded from the assessment. This is because I have shown in earlier work that even though frequently and extensively used, the status of single Vietnamese kin terms in otherwise English discourse remains ambiguous due to conflicting evidence (Nguyen, 2018).

Type	N	Frequent ( $\geq 10$ )	Diffuse ( $\geq 5$ )	Type	N	Frequent ( $\geq 10$ )	Diffuse ( $\geq 5$ )
okay	48	✓	✓	chef	21	✓	×
yeah	31	✓	✓	copy	12	✓	×
homework	25	✓	✓	comment	11	✓	×
cent	19	✓	✓	exam	11	✓	×
game	15	✓	✓	<i>cái</i> (CLS)	9	×	✓
lecture	15	✓	✓	<i>thì</i> (CONJ)	9	×	✓
gym	13	✓	✓	book	8	×	✓
<i>hả</i> (DM)	11	✓	✓	lunch	8	×	✓
oh	11	✓	✓	test	5	×	✓

Table 3.2: Frequent and widespread single other-language items in CanVEC

It is apparent from Table 3.2 that most tokens listed are single English words (N=15/18), and that discourse markers (DM) such as ‘okay,’ ‘yeah,’ *hả*, ‘oh’<sup>39</sup> are particularly prevalent, accounting for almost half of both frequent and widespread items (N=4/9). Though it has often been assumed that DMs are language-neutral and generally serve the same discourse function in both bilinguals’ languages, this is not always the case (Balukas & Koops, 2015). Data from CanVEC suggests that some DMs are language-specific and never occur in other-language context. For example, ‘well,’ ‘you know,’ ‘I bet,’ ‘I guess’ only occur in an English environment, while *ồ* ‘oh,’ *ờ* *thì* ‘uhm well,’ *thật hả* ‘really?’, *vâng* ‘yeah’ occur only in a Vietnamese context. As for those that occur in both, only items listed as recurrent **and** widespread (left half, Table 3.2) are marked as established borrowing. These items are subsequently language-tagged as part of their surrounding discourse. The remainder of this group includes ‘uh,’ ‘ha,’ ‘ah,’ which are neither frequent nor diffuse and therefore marked as language-neutral. Overall, given that DMs are either considered language-neutral or tagged as their surrounding language, clauses containing DMs as single other-language insertions are in any case monolingual clauses.

### 3.3.2.3 Automatic translation

Having verified clause-level language tags, we next automatically translated all the Vietnamese and mixed clauses using the Google Translate API.<sup>40</sup> Although we could have segmented and translated only the Vietnamese subsequences in the mixed clauses (as we did for tokenisation

<sup>39</sup>All English-spelt DMs were pronounced in English, and Vietnamese-spelt DMs in Vietnamese. For example, ‘oh’ represents the English pronunciation /əʊ/, while Vietnamese *ồ* ‘oh’ represents the Vietnamese pronunciation /o-ɿ/. As we will later see, while ‘oh’ occurs in both Vietnamese and English contexts, *ồ* is strictly found in Vietnamese clauses only.

<sup>40</sup><https://cloud.google.com/translate/>

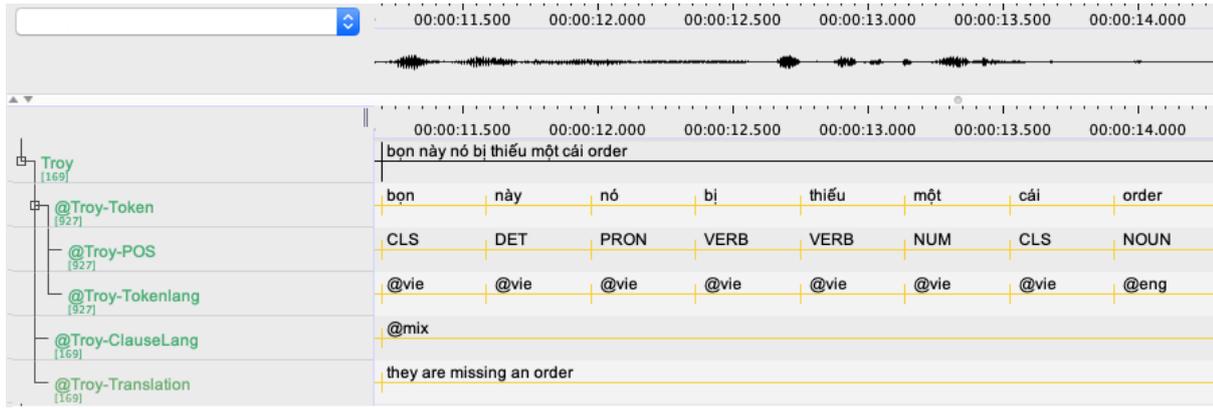


Figure 3.2: A sample speech-tier alignment in ELAN

and POS-tagging), we instead sent the entire bilingual clause to the translation API this time. This is because machine translation systems are usually designed to handle unknown words and also tend to perform better on longer sequences of input (for more context), and so we expected better translations at the clause level rather than the sub-clause level.

All the output was then imported back into ELAN and distributed across various tiers. [Figure 3.2](#) hence shows how a transcribed, time-aligned clause for each speaker is associated with separate sub-tiers for tokens, token POS-tags, token language tags, and a clause language tag. This link between transcription, encoding and speech signal not only assists with data transparency, but also facilitates preliminary analysis.

[Table 3.3](#) provides some basic statistics concerning the overall composition of CanVEC after automatic annotation.

Type	Clauses	Tokens
Vietnamese	7,508	45,640
English	2,582	15,523
Mixed	2,721	22,094
Non (X & non-clausal fragments)	1,236	3,462
<b>TOTAL</b>	<b>14,047</b>	<b>86,719</b>

Table 3.3: Basic linguistic statistics of the CanVEC corpus

### 3.3.3 Evaluation

#### 3.3.3.1 Language marking and POS-tagging

Recall that one of the major aims of this dissertation is to provide the first high-quality, reliable Vietnamese-English natural speech corpus that is accessible for future research. Now that

we have systematically language-marked and POS-tagged the data, the next step is to evaluate the performance of our method. For automatic POS-tagging and language identification, 100 clauses of each type (i.e. monolingual Vietnamese (N=7,508), monolingual English (N=2,582), and mixed (N=2,721)) were randomly selected and manually assessed in terms of label accuracy. Equation 3.1 shows how accuracy was calculated, in which X is a specific level type, including token language tags, token POS-tags, and clause language tags.

$$\text{Accuracy (\%)} = \frac{\# \text{ Correct labels of X}}{\# \text{ Total labels of X}} \times 100 \quad (3.1)$$

Applying this, Table 3.4 reports the results for each level of annotation.

Type	Token Language	Token POS	Clause Language
Vietnamese	96%	76%	99%
English	100%	99%	100%
Mixed	97%	75%	99%

Table 3.4: Accuracy report for each level of semi-automatic annotation

It is apparent from the results that while language identification was almost perfect at both the token level and the clause level, most likely because Vietnamese and English words tend to be orthographically distinct (and appropriately represented by the selected transcription convention), POS-tag results for Vietnamese were noticeably less robust. This is likely because Vietnamese POS-taggers are not only typically trained on less data than English POS-taggers, but they are also unlikely to be well-suited to speech data. Specifically, spoken Vietnamese is significantly different from written Vietnamese in that the spoken variety is characterised by extensive use of discourse markers and lexical variation due to regional dialects. This means text-trained POS-taggers are not always optimal for analysing spoken discourse, particularly in low-resource languages.

Additionally, results for mixed-clause POS-tags were also lower compared to English, although this is most likely for the same reason that the results for Vietnamese POS-tags were low. Alternatively, since mixed clauses were split into smaller subsequences before being sent to the appropriate monolingual tagger, it might also be the case that the sequences were too short to give the tagger enough context to assign a reliable tag. It should be noted that for mixed clauses, 100% of POS-tag errors are Vietnamese POS tags (N=56).

Further error analyses also showed that the majority of Vietnamese POS-tagger errors involved pronouns and classifiers (83%, N=162/181). Pronoun-wise, this is very likely due to the complex system of Vietnamese personal reference, which uses different pronouns, kin terms, NPs (including NPs that consist of only a CLS and a DET without an overt N) and personal

names in different contexts (Nguyen, 2018). Crucially, while kin terms and personal names are frequently used as personal pronouns in spoken discourse, they are fairly unproductive in written news texts or narrative. Given that Vietnamese POS-taggers are trained using written resources, they understandably struggle with the spoken domain where a different set of pronouns is dominant. Table 3.5 lists the distribution and proportion of these errors across the evaluation set. Equation 3.1 is again applied here, with X now being defined as a specific type of POS label (PRON, CLS, N, etc.)

Correct tag	Tagged as	N	%
Pronoun (PRON)	Classifier (CLS)	59	63%
	Noun (NOUN)	26	28%
	Particle (PRT)	7	7%
	Preposition (PREP)	2	2%
Classifier (CLS)	Pronoun (PRON)	56	82%
	Interjection (INTJ)	12	18%

Table 3.5: The distribution of PRON and CLS POS-tag errors (N=162 errors/100 sample Vietnamese clauses and 100 mixed clauses)

As we can see from the results in Table 3.5, PRON and CLS were most frequently mistaken for each other (63% and 82%) rather than for something else. Linguistically, this is highly likely due to the fact that one of the most frequent kin-terms-used-as-pronoun (i.e. *con* ‘child’) is homophonous with the Vietnamese general animate classifier. However, this is arguably a positive error as it shows that the negative effect was confined only to a limited domain (i.e. PRON and CLS) and not quite spread out to other types.

Importantly, despite the difficulties with PRON and CLS, results for other Vietnamese POS-tags, particularly Nouns (NOUN)<sup>41</sup>, Verbs (VERB), Adverbs (ADV), and Prepositions (PREP) remain particularly strong, with error rates in the range of 1-5% (see Appendix H for full confusion matrices). This means that, barring PRON and CLS, other Vietnamese POS-tags can be reliably extracted from the corpus.<sup>42</sup>

<sup>41</sup>Note that although it is apparent from Table 3.5 that 1/3 of PRON were incorrectly tagged as N, these only count towards PRON errors and do not count towards N error rates. This is because in Vietnamese (and many other languages), PRON is considered an open-class subset of N, and hence a PRON can be an N in essence, but not vice versa.

<sup>42</sup>At the time of writing, the potentially problematic POS-tags in the corpus have not been fixed. This is a task for future work. As the following chapters will show, however, this does not impact the analysis. Due to the linguistic nature of the variables of interest, none of these forms is extracted based on POS tags but rather by lexical forms (English) and manual retrieval (some English, all of code-switching and Vietnamese).

### 3.3.3.2 Translation

To evaluate the quality of the automatically translated clauses, I randomly selected 100 monolingual Vietnamese and 100 mixed clauses and rated them in terms of commonly used qualitative metrics, namely fluency, comprehensibility, and semantic adequacy (Koehn, 2009; Dorr, Snover & Madnani, 2011). Each of these metrics is defined as follows:

- Fluency: Does the translation sound natural in the target language (i.e. English)?
- Comprehensibility: Does the translation make sense on its own, independently of the source clause? If yes;
- Semantic adequacy: Does the translation retain the intended meaning in the source clause?

I am aware that these metrics overlap to some extent<sup>43</sup>, but there is no straightforward solution to this problem. In fact, robust machine-translation evaluation is still an active area of research; and although lots of different metrics exist (e.g. Papineni, Roukos, Ward & Zhu, 2002; Snover, Dorr, Schwartz, Micciulla & Makhoul, 2006; Lavie & Agarwal, 2007; Lo & Wu, 2013), no metric is perfect. Formalising the factors that determine the quality of a translation is still a hard task (see e.g. Moorkens, Castilho, Gaspari & Doherty, 2018 for an overview), and it is worth stating that the goal of this evaluation is not to formally evaluate the performance of the Google Translate API on Vietnamese and code-switching speech, but rather to ascertain the quality of the automatic translations for reasons of corpus reliability.

With this in mind, I then assigned a binary Yes/No judgement for each metric to each clause in the sample. A binary scale rather than a Likert scale was used because clauses were short enough to expect fewer mistakes from the translation system (Koehn, 2009, p.218). Results are reported in [Table 3.6](#).

As the results illustrate, the overall quality of corpus translation for monolingual Vietnamese is relatively positive, with more than half of the clauses meeting all three requirements of semantic adequacy, fluency, and comprehensibility. Additionally, it is worth noting that machine translation performed best at comprehensibility on both sets of data, scoring 80% and 72% on monolingual Vietnamese and mixed clauses respectively. Although maintaining fluency in mixed clauses still seems to be a particular area of difficulty (54%), the fact that a majority of translations were rated fluent, comprehensible, and semantically adequate suggests that the output is still reasonably useful to users of CanVEC.

---

<sup>43</sup>No sentence is incomprehensible but semantically adequate, and hence if the sentence is marked 'Not comprehensible,' it is also marked as 'Semantically inadequate.'

Metric	Vietnamese	Mixed	N Metrics Satisfied	Vietnamese	Mixed
Fluent	77%	54%	0	11%	20%
Comprehensible	80%	72%	1	11%	14%
Semantically adequate	67%	64%	2	22%	23%
			3	56%	43%

(a) The proportion of clauses meeting each criterion per metric

(b) The distribution of clauses meeting at least N criteria

Table 3.6: An overview of the translation quality in a sample of 100 Vietnamese and 100 Mixed clauses in CanVEC

In terms of specific errors, I found that similar to Vietnamese POS taggers, machine translation seems to struggle most with Vietnamese pronouns. Example (17) illustrates contrasting occasions when the pronoun was translated incorrectly and correctly in a monolingual Vietnamese and mixed clause respectively. In particular, the first person subject *con* (kin term meaning ‘child’) was erroneously translated as a 3SG common noun in the monolingual Vietnamese clause, but accurately translated as a 1SG subject pronoun in the mixed clause.

- (17) a. **Input:** *con đi bộ.* [Monolingual Vietnamese]  
**Gloss:** 1SG.kin go foot  
**Machine translation:** child walking.  
**Human translation:** I walked.

(Penny.Marie.Rory.0912, 11:48.8–11:49.4)

- b. **Input:** *mà giống-như con* pick up a little bit of Busan Busan dialect.  
**Gloss:** but like 1SG.kin [Mixed Vietnamese-English]  
**Machine translation:** but like I pick up a little bit of Busan Busan dialect.  
**Human translation:** but like I pick up a little bit of Busan Busan dialect.

(Tim.Jess.Chloe.0705, 08:03.7–08:10.2)

Although this is only an isolated example in the evaluation sample, it is nevertheless surprising that the correct translation is found in a mixed clause, which typically scores lower in the evaluation overall. This observation leads me to suspect that the better-resource participating language in code-switching (i.e. English in this case) possibly contributes to enhancing the accuracy of machine translation in processing the lower-resource language (i.e. Vietnamese). However, as I do not have a large enough sample of data to further probe this, future experiments are needed to appropriately test this hypothesis.

### 3.3.4 Summary

Overall, our method for semi-automatically annotating CanVEC data represents an opportunity to overcome the traditionally expensive process of manual annotation. Here, the system is simple and effective enough that it can be extended to processing other language pairs, especially those involving a low-resource, minority language.<sup>44</sup> Although this is not always straightforward and some tasks remain challenging, the overall performance is good enough to render the output utilisable. I provide an example of an annotated continuous dialogue in CanVEC in [Appendix I](#).

## 3.4 Chapter summary

In this chapter, I described the foundation and the core elements of CanVEC, an originally built, systematically annotated spontaneous speech corpus of the Vietnamese bilingual community in Canberra. In doing so, the chapter recognised the need for, and proposed a method of, annotating a mixed-language corpus, which can considerably speed up the creation of new corpora in future research. The time-aligned CanVEC corpus materialised as a result, consisting of 45 speakers of two generations, 10 hours of spontaneous speech, and approximately 90,000 words. Aside from serving as the empirical foundation for the rest of the discussion in this dissertation, it also makes available to future research the first digitalised, comparative Vietnamese-English data of the Canberra Vietnamese vernacular.<sup>45</sup>

---

<sup>44</sup>Thanks to the financial support from the Cambridge Language Sciences Incubator Fund, the method has been extended to Hindi-English. Results are promising, with a reported accuracy rate of 90.68% for language tagging. See Kidwai, Bryant, Nguyen & Biberauer (2019) for further details.

<sup>45</sup><https://github.com/Bak3rLi/CanVEC>.



## **Part II**

# **Cross-generational variation in the Vietnamese heritage language of the Canberra Vietnamese community**



---

# THE MATRIX LANGUAGE IN THE COMMUNITY

---

## 4.1 Introduction

In this part, I put CanVEC to use and investigate cross-generational linguistic variation in the Canberra Vietnamese-English bilingual community. The first theoretical perspective to be considered is the Matrix Language Turnover Hypothesis (Myers-Scotton, 1998), one of the most well-known, but rarely tested, hypotheses in relation to cross-generational language variation and shift. The hypothesis refers to a situation in which the original Matrix Language (ML), i.e. the language that provides the morphosyntactic frame for a bilingual complementiser phrase (CP), becomes, for many speakers in a given community, the Embedded Language (EL), i.e. the language that is merely ‘inserted’ into the structural frame provided the ML, and vice versa. In most cases, the original ML is the minority language (i.e. the language with less socio-political power), whereas the new ML is the language of the majority (i.e. the language with more socio-political power). Due to higher prestige and/or greater socio-economic and political power, the majority language then takes over and replaces the minority language as the ML for most bilingual CPs produced by community speakers.

In the context of the Canberra Vietnamese community, given that English has always been the language with greater socio-economic and political power ([Chapter 2](#)), we might expect that an ML Turnover would take place in that direction; i.e. that English would replace Vietnamese as the ML in bilingual CPs. As Myers-Scotton (1998) argues, when such a ‘turnover’ is complete, language shift or language death is likely to follow. Studying a ‘ML turnover,’ then, is potentially illuminating in capturing ongoing changes within the community and envisioning the future of a heritage language. What I aim to achieve in this chapter is thus to probe Vietnamese heritage language ‘indirectly’ by investigating its participation in the bilingual code-switching subset of the corpus.

Another motivation for adopting the ML Turnover Hypothesis is an opportunity to test the Matrix Language Framework (MLF) within which it is embedded. Specifically, in order to explore whether an ML Turnover is complete or underway, an ML for each bilingual clause in CanVEC (N=2,721) needs to be determined using the MLF principles (Myers-Scotton, 1993, 2002; see also [Chapter 1](#)). Although the MLF is one of the most influential models in language contact, its support has mainly come from language pairs that are typologically different in terms of their clausal word order, or else have vastly different inventories of inflectional morphology. Vietnamese-English as an under-described language pair in which both participating languages are SVO and morphologically limited has never been tested. While the linguistic nature of this language pair is thus already a concern regarding the applicability of the MLF, what this chapter aims to achieve is to test—using new and empirical data—how far we can take the ‘universal’ theoretical assumptions of the MLF.

The structure of the chapter is as follows. It first describes the ML Turnover Hypothesis and its associated model, the MLF (§4.2), and introduces two basic principles: The Morpheme Order Principle and the System Morpheme Principle which are used to identify the ML of the clause. It next continues with a review of previous work that has made use of the MLF model and the ML Turnover Hypothesis (§4.3), before applying it to the CanVEC dataset (§4.4). The results highlight several limitations of the MLF model, specifically calling into question again the assumption of a monolingual baseline (see [Chapter 1](#), §1.1) and the notion of a ‘Composite ML’ (§4.5). Finally, I evaluate whether an ML Turnover is present (§4.6), the direction the change is heading in, and what conclusions we can draw from it (§4.7).

## 4.2 The Matrix Language and Matrix Language Turnover

### 4.2.1 Myers-Scotton’s Matrix Language Framework (MLF)

One of the most prominent views in the current literature is that there exists an asymmetrical relationship between the two languages in any bilingual discourse. This idea was first put forward by Joshi (1985), stipulating that in any given combination of two languages, the structural contributions of the languages are not equal. As Joshi argued, despite systematic interactions between the languages which may sometimes give rise to mixed utterances, ‘speakers and hearers generally agree on which language the mixed sentence is ‘coming from’ (Joshi, 1985, pp.190–191),

thereby selecting one ‘main language’ for the utterance.<sup>46</sup> It is this main language that provides the morphosyntactic frame (i.e. the ‘Matrix Language’ (ML) in Myers-Scotton’s term), while the other language is merely inserted into the ML’s pre-existing structure (hence the ‘Embedded Language’ (EL)). While the ML can provide all kinds of grammatical categories, switches to the EL are restricted to open categories only, such as nouns or lexical verbs.

Building on this notion of asymmetry, Myers-Scotton (1993) was the first to formalise these ideas into what later became known as the Matrix Language Framework (MLF), which has since enjoyed considerable research attention. In essence, the MLF centres itself around three core principles: the **Matrix Language Principle**, the **Asymmetrical Principle** and the **Uniform Structure Principle**. Respectively, these principles specify that:

- (i) **Matrix Language Principle:** only one language supplies morphosyntactic structure for any given mixed clause in which two languages are combined (the ML);
- (ii) **Asymmetrical Principle:** the ML is unambiguously identifiable in these clauses; and
- (iii) **Uniform Structure Principle:** all structural elements (i.e. functional morphemes) are preferentially from the ML rather than the EL in order to maintain well-formedness.

In what follows, I will describe the key arguments of the MLF.

#### 4.2.1.1 The Content-System Morpheme distinction

One fundamental aspect of the MLF model is the distinction between content morphemes and system morphemes. In the earliest versions of the MLF, Myers-Scotton (1993, 1997) proposed [quantification] as a feature to distinguish these two types of morphemes: those that have a [+quantification] feature such as quantifiers, specifiers, or inflectional morphology are system morphemes, while those that do not are content morphemes. In her later work, however, content morphemes are defined as those that can assign or receive thematic roles, with semantic and pragmatic features, while system morphemes ‘largely indicate relations between content morphemes’ (Myers-Scotton, 2002, p.15). Accordingly, open-class words such as nouns, verbs, adjectives are straightforwardly considered content morphemes, and closed-class function words such as determiners, number/gender/case marking, and prepositions fall into the class of ‘system morphemes.’ This distinction is further developed into what she terms the ‘4M model,’ which is summarised in [Figure 4.1](#).

<sup>46</sup>It should be noted that bilinguals do not necessarily always mix languages or engage in code-switching (see e.g. Bullock & Toribio, 2009; Gardner-Chloros, 2009 for some helpful overviews). Such mixing is, however, a community norm in the Canberra Vietnamese bilingual community, as previously described in [Chapter 2](#), §2.4.4 and [Chapter 3](#), [Table 3.3](#).)

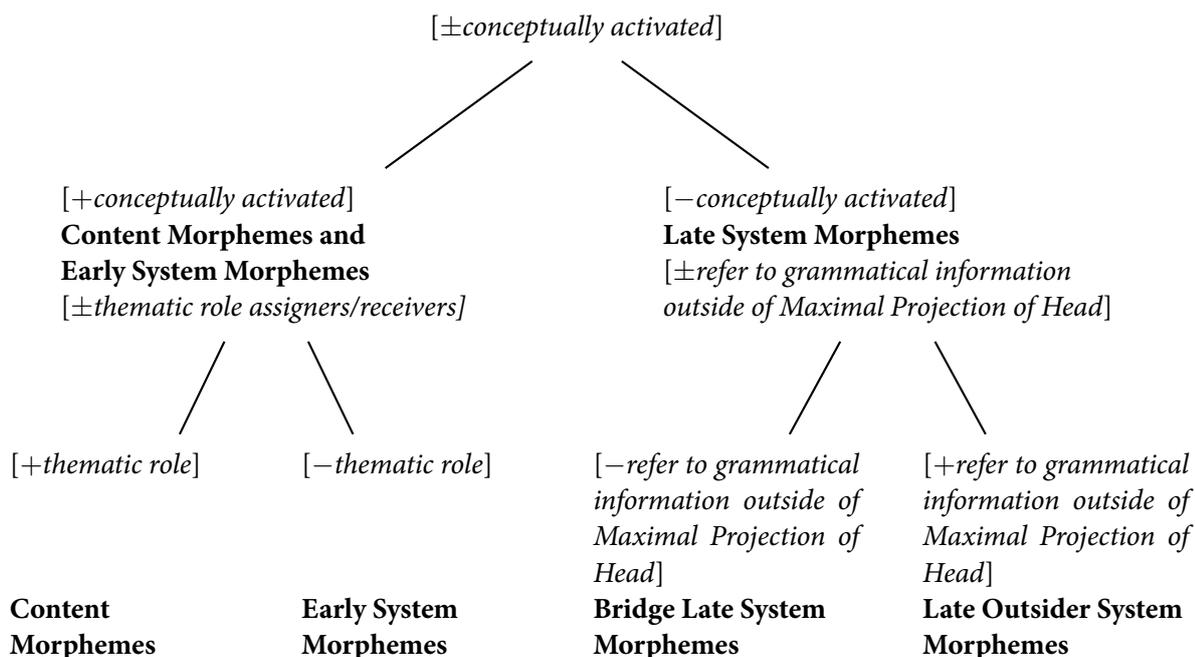


Figure 4.1: Different types of morphemes in the 4M model, Myers-Scotton (2002, p.73)

In this model, Early System Morphemes are conceptually activated and index ‘semantic and pragmatic meanings that satisfy speakers’ intention’ (2000b, p.1055) (e.g. determiners and plural markers in English). Late System Morphemes, by contrast, do not require activation at the lemma level (e.g. verbal agreement). Bridge Morphemes, such as possessive ‘of’ and ‘s’ in English, also belong to this category and are characterised by their ability to ‘unite morphemes into larger constituents’ (Myers-Scotton & Jake, 2000a, p.4). Nonetheless, as we will shortly see, the most important type of system morphemes that are central to the identification of the ML are Late Outsider System Morphemes. Following Myers-Scotton & Jake (2000a), Late Outsider System Morphemes are defined as those depending on information outside their immediate maximal projection (i.e. an XP of some kind) for their forms. Typical Late Outsider Morphemes are subject-verb agreement affixes and case affixes. Example (18) below provides an illustration.

- (18) Tanner<sub>1</sub>: *nó* bites *cái* fingers *cúa* *con*.  
           3SG      CLS      of 2SG.kin  
           ‘It bites your fingers.’

(Tanner.Nina.0609, 07:14.7–07:18.3)

In this example, the subject-verb agreement ‘-s’ attached to the content morpheme ‘bite’ is considered a Late Outsider Morpheme as its presence is conditioned by checking information (i.e. the subject *nó*) outside its immediate maximal projection (i.e. a VP in this case).<sup>47</sup> In contrast,

<sup>47</sup>Note that subject-verb agreement ‘-s’ can also be said to start outside the VP. According to Government and Binding theory for example, the agreement element ‘-s’ would start under INFL/T agreeing with the lexical verb. This is to say that ‘-s’ is not inherently connected to the VP, but under the MLF model, it is considered part of the VP agreeing with the subject, which is outside its maximal constituent.

the plural ‘-s’ attached to ‘finger’ is an Early System Morpheme as it is indirectly selected by the head content morpheme ‘finger.’ Specifically, this plural marking ‘-s’ serves to add the concept of number, thereby completing the semantic features of the speakers’ intentions. Contrary to all Late System Morphemes, Early System Morphemes are indexed within the same maximal projection of the head content morphemes that select them. Finally, the Vietnamese possessive marker *của* is considered a bridge Late System Morpheme as it is responsible for joining ‘*cái* fingers’ with the 2SG kin term *con* to create a larger possessive constituent.<sup>48</sup>

#### 4.2.1.2 The Matrix Language-Embedded Language distinction

Another key distinction in the MLF model is that of the Matrix Language (ML) and the Embedded Language (EL). According to Myers-Scotton (2002), the ML is the language that contributes more by supplying the grammatical structure for a mixed utterance, while the EL is only responsible for inserted materials within the ML frame. In other words, the ML sets the morphosyntactic frame for the utterance, and the EL only works to ‘fill in the gap.’ To identify the ML, Myers-Scotton proposed two universal principles that can be applied to any language pairs:

1. *The System Morpheme Principle:*

In ML+EL constituents, all system morphemes which have grammatical relations external to their head constituent (i.e. which participate in the sentence’s thematic role grid) will come from the ML.

2. *The Morpheme Order Principle:*

In ML+EL constituents consisting of singly occurring EL lexemes and any number of ML morphemes, the surface morpheme order (reflecting surface syntactic relations) will be that of the ML.

(Myers-Scotton, 2002, p.59)

In Myers-Scotton’s model, both principles are applied simultaneously, meaning both principles must be satisfied in any given code-switching (CS) clause. Essentially, the ML is supposed to be the language determining the word order and the language that supplies function words. The ‘constituents’ referred to in the above principles are either an ML+EL mixed constituent, an EL island containing only EL morphemes, or an ML island containing only ML morphemes. A

<sup>48</sup>As I later discuss in §4.3.1.2 and §4.6.2, the Vietnamese classifier *cái* also fits the definition of an Early System Morpheme, mainly on the basis that its form changes depending on the features of the head noun within its maximal projection.



as the unmarked choice—just because it gives preference to keeping structure uniform across the CP. In this sense, barring the exception of EL islands (e.g. *la semaine* ‘the week’ and *les permis* ‘the permits’ in example 21), the MLF posits that all system morphemes are sourced from the ML.

### 4.2.2 The Matrix Language Turnover Hypothesis

In contact linguistics, it is often believed that lexical borrowing is the beginning of language contact, followed by CS and bilingualism (Thomason & Kaufman, 1998). Where contact between two languages occurs, however, it is not unusual that one of the languages has greater socio-economic or political power, and will either gradually or rapidly become more dominant in the speech community. As Myers-Scotton (2002, p.52) suggests, ‘there is always a power differential between the languages involved—simply because access to sources of power (e.g. high-level jobs, educational facilities, or governmental services) are not equally distributed.’ In this case, the language of the minority is naturally the one that suffers and starts losing its place in favour of the language that offers more socio-economic benefits. In an immigration setting, this often means that the language of the host society will eventually take over, and, by the second or third generation, become the major medium of communication in the community (Alba, Logan, Lutz & Stults, 2002; Sofu, 2009; Habtoor, 2012).

To capture such linguistic outcomes of language contact, Myers-Scotton (1998) introduced the Matrix Language Turnover Hypothesis, which stipulates that in communities where there is widespread CS or convergence within a CP and where there is ‘a dramatic shake-up in the socio-political balance’ in favour of the prestigious language, an ML Turnover will result (p.300). An ML turnover is defined as a situation in which ‘the main language which had structured constituents becomes the structurally minor (i.e. the Embedded Language (EL)); in turn, the language which had been the minor language regarding structure becomes the ML’ (Myers-Scotton, 1998, p.299). In other words, the original ML responsible for setting the morphosyntactic frame in bilingual CPs becomes the EL supplying content morphemes, and vice versa. Myers-Scotton goes on to argue that it is this turnover of the ML that sets the stage for structural change. The ML Turnover Hypothesis is summarised in [Figure 4.2](#).

As [Figure 4.2](#) illustrates, widespread intra-sentential CS or convergence, or both, are taken as necessary conditions for the incursion of one language into another. For the purpose of the hypothesis, CS is defined as ‘the use of morphemes from two or more linguistic varieties in the same CP’ whereas convergence is ‘the use of morphemes from a single linguistic variety, but with parts of their lexical structure coming from another source’ (Myers-Scotton, 1998, p.291). A way

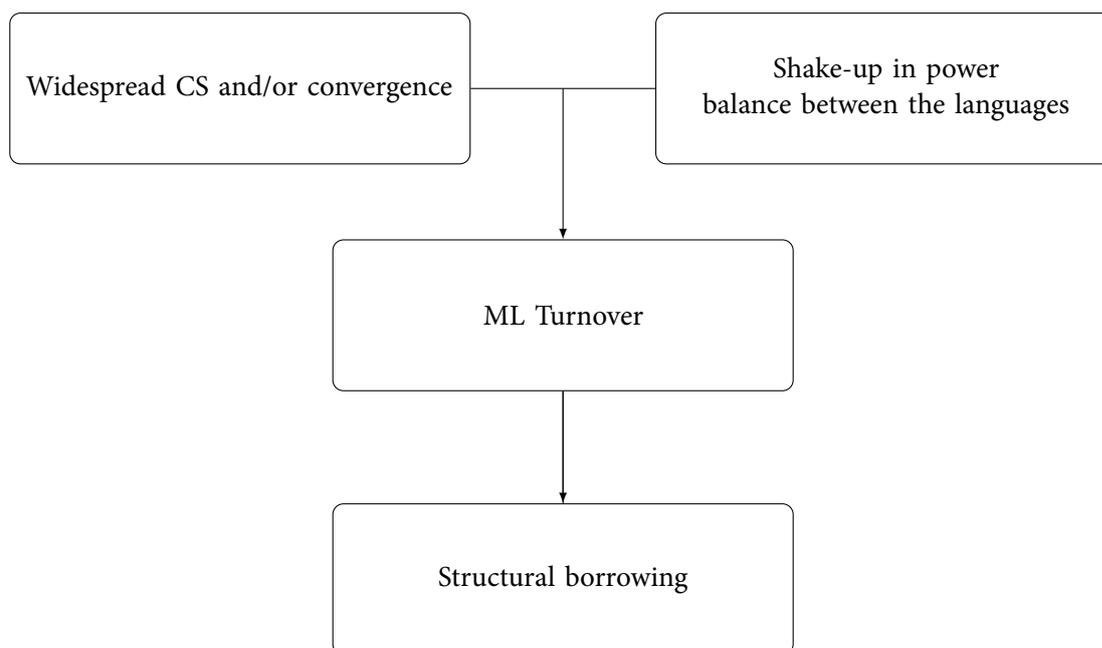


Figure 4.2: The ML Turnover Hypothesis

to falsify the ML Turnover Hypothesis, as Myers-Scotton herself recognises, ‘would be to show that structural borrowing occurs where these conditions are not present’ (1998, p.300).

Accordingly, three possible scenarios for an ML Turnover are then proposed:

1. **Arrested ML Turnover:** The old ML (i.e. the minority language) remains the main medium of communication within the minority community, with some degree of structural borrowing from the majority language. As Myers-Scotton describes it, at this stage the old ML’s content morphemes are used alongside the new ML’s system morphemes, some of the new system morphemes are reproduced in the old ML. Extensive CS is not necessarily still present. Examples of this scenario could be innovation or borrowing a new linguistic element from the EL. In the context of Vietnamese-English for example, ‘structural borrowing’ could thus involve the use of English definite articles in Vietnamese utterances. It is important to note that, by definition, structural change here is confined to morphology, i.e. the borrowing/distribution of system morphemes from the new ML.
2. **Composite ML:** This scenario refers to a situation when the morphosyntactic frame of a CP comes from both participating varieties, and this ‘Composite ML’ fossilises as the main medium of communication. This is hypothesised as the stage where two languages converge, splitting and recombining abstract lexical structure. This convergence is explained under the Abstract Level model, which posits three abstract levels of the lemma in which convergence can manifest: the **lexical-conceptual level**, where ‘language-specific semantic/pragmatic feature bundles’ are activated (Myers-Scotton & Jake, 1995, p.987);

the **predicate-argument level**, where the thematic structure is mapped onto grammatical relations; and the **morphological-realisation level**, where grammatical relations are realised on the surface (e.g. agreement, word order, case marking, etc.). Myers-Scotton (1998, p.301) attributes this to most ‘split languages’ that ‘largely show the grammar of one language and the lexicon of another’ (see Myers-Scotton, 2003 for further discussion on this).<sup>50</sup>

3. **Complete Turnover:** A situation when a turnover goes to ‘completion,’ and is taken as ‘the most common outcome’ of languages in contact (Myers-Scotton, 1998, p.294). This is when language shift follows, CS falls away and the ‘Composite ML’ is replaced by a single variety of the new ML during CS.

The idea of a Composite ML, however, has been particularly subject to criticisms (e.g. Boussofara-Omar, 2003; Auer & Muhamedova, 2005; Gardner-Chloros, 2009). First and foremost, a ‘Composite ML’ by its nature defeats the fundamental MLF idea that there is always one dominant language in bilingual discourse. Specifically, if we accept that more than one variety can participate in setting out the grammatical structure of the ML+EL constituents, then the ML-EL hierarchy and the system-content morpheme distinction are substantially weakened (Boussofara-Omar, 2003). Second, if we have utterances that show the lexicons from one language but the grammar of another, then by definition, the language of the grammatical structure is already the ML. It is thus unclear why the notion of a ‘Composite ML’ is needed, and what explanatory power it actually holds.

It is worth recognising, however, that the ML Turnover Hypothesis is still highly relevant to Vietnamese-English in any case. Recall from [Chapter 2](#) that the Vietnamese community in Canberra has long been a minority community. English, as the only medium of education and employment, consequently enjoys higher socio-economic status and offers greater employment opportunities than Vietnamese. This, together with the lack of a designated, clustered ‘Vietnamese area’ in Canberra, an increasing number of second-generation speakers, and the high English proficiency of first-generation speakers in the community, give us reasons to expect that Vietnamese is becoming less pertinent. Additionally, the considerable degree of intra-CP CS as we have previously seen in [Chapter 3](#) ([Table 3.3](#)) could also suggest a probable turnover in the

<sup>50</sup>Note, however, that not all mixed languages show ‘the grammar of one language and lexicon of another.’ Michif (the language of the Métis people of Canada and the U.S), for example, combines Cree and Métis French and exhibits clear splits within their grammatical system. Specifically, while all the nominal elements such as lexical gender and adjective agreements in Michif derive from Métis French, its clausal and verbal elements are taken from a southern variety of Plains Cree (which a dialect of Cree). See Bakker, 1997 et seq. for a comprehensive overview of this language.

ML according to the ML Turnover Hypothesis, which potentially sets the scene for structural change to follow.

Before testing the ML Turnover Hypothesis, however, it is first important to understand what has previously been learned about the MLF model and the ML Turnover Hypothesis.

## 4.3 Application of the MLF in the literature

### 4.3.1 Previous work using the MLF

As one of the most prominent models in language contact, the MLF has enjoyed enormous attention from those taking a structural approach to CS. Myers-Scotton (2006, p.248) has claimed that a range of studies demonstrated ‘the universality of support’ for the model, ‘no matter which languages are involved.’ In fact, the MLF has been successfully applied to various sets of data, including, but not limited to, Swahili-English (Myers-Scotton, 1993); Zulu-English and Sotho-English (Finlayson, Calteaux & Myers-Scotton, 1998); German-English (Fuller & Lehnert, 2000), Welsh-English (Deuchar, 2006; Deuchar et al., 2018), and Igbo-English (Ihemere, 2016, 2017). Instead of reviewing all of these studies, I will consider here only those aspects of previous work that are of particular relevance, namely the predictive power of the MLF (§4.3.1.1) and its application in language pairs with limited morphology and homologous word order (§4.3.1.2).

#### 4.3.1.1 Predictive power of the MLF

Despite being used as a productive platform to analyse CS data, the predictive power of the MLF is still debatable. For example, in a study focusing on Arabic diglossic switching between Tunisian Arabic (TA), a dialectal variety of Arabic spoken in Tunisia, and Fushaa, a blanket term for Classical Arabic/Modern Standard Arabic, Boussofara-Omar (2003) put forward counterexamples to the two main principles of the MLF (i.e. the **System Morpheme Principle** and the **Morpheme Order Principle**), showing that:

- (i) it is possible for both participating languages to contribute system morphemes to the clause; and
- (ii) there exist cases where all morphemes come from Fushaa, but word order reflects that of Tunisian Arabic.

In (22), for instance, Boussofara-Omar suggests that the system morphemes come from both Tunisian Arabic (i.e. discontinuous negation marker *ma-...f*) and Fushaa (i.e. first-person singular imperfect marker *a- ʕ*). It is important to note that while this example was more of a challenge to an earlier version of the System Morpheme Principle<sup>51</sup> of the MLF, Boussofara-Omar's claim that both varieties participate grammatically in the 'framing' of the CP remains valid here. As Boussofara-Omar further notes, the constituent *ma-ʔ a- ʕ taqid-f* 'I don't believe' is not an Embedded Language island (since it contains system morphemes from both varieties), and therefore cannot be taken as a lawful violation of the MLF (§4.2.1.2).

- (22) *ma-ʔ a- ʕ taqid-f* [Tunisian Arabic-Fushaa]  
 NEG-1SG IMP-believe-NEG  
 TA-F F TA  
 'I don't believe.'  
 (TA = Tunisian Arabic, F = Fushaa, Boussofara-Omar, 2003, p.39)

Similarly, in example (23a), Boussofara-Omar suggests that while all morphemes are from Fushaa, the structure [because + VP + NP] is that of the dialectal pattern of TA (as in (23b)), rather than the pattern [because + NP + VP] expected of Fushaa (as in (23c)). In other words, while all system morphemes come from one variety, word order reflects grammar of the other.

- (23) a. *liʔ anna laa ya-nhaḍ al-ʔ adab* [...] [Actual utterance]  
 because NEG 3MAS-SG-IMP-advance DEF-literature  
 F F F F  
 'Because literary production does not progress.. [...]'  
 b. *ʕ la xaatar ma yi-tqaddam il-ʔ adab* [TA word order]  
 because NEG 3MAS SG [TA subordinator + VP + NP]  
 'Because literary production does not progress...'  
 c. *liʔ anna al-ʔ adab-(a) laa ya-nhaḍ* [Fushaa word order]  
 because DEF-literature-(ACC) NEG 3MASC-SG-IMP-advance [F subordinator + NP + VP]  
 F F F F  
 'Because literary production does not progress...'  
 (F = Fushaa, Boussofara-Omar, 2003, p.42)

Responding to these examples, Myers-Scotton (2004) claims that 'the MLF model was formulated to cover CS between language varieties that are separate languages (i.e. not mutually intelligible varieties, such as dialects)' (p.89). However, as Wang (2007) points out, this defence is problematic in at least two respects. First of all, the boundary between a language and a dialect

<sup>51</sup>In the earliest version of the MLF in 1993, the System Morpheme Principle stated that 'Within ML+EL constituents, **all active system morphemes** are from only one of the languages participating in CS, i.e. the ML' (Myers-Scotton, 1993, p.83). In later work, 'all active system morphemes' was reformulated as 'all system morphemes that have grammatical relations external to their head constituent'; i.e. those that participate in the sentence's thematic role grid (Myers-Scotton, 2002). The requirement for 'all system morphemes' to be sourced from the ML, however, remains an active premise of the MLF. In particular, the current 'Uniform Structure Principle' of the model (briefly mentioned in §4.2.1, §4.2.1.2, and later discussed in §4.3.1.2) stipulates that all system morphemes are expected to come from the ML, 'just because it gives preference to keeping structure uniform across the CP' (Myers-Scotton, 2002, p.120).



power of the MLF model. In fact, as we will shortly see, the MLF model makes explicit predictions about the determiners being sourced from the ML.

In one of the first studies directly juxtaposing the MP and the MLF predictions, Herring, Deuchar, Parafita Couto & Moro (2010) test the MLF vs. the MP accounts for the determiner phrase in Welsh-English (using the *Siarad* corpus<sup>53</sup>) and Spanish-English (using the *Miami* corpus<sup>54</sup>). In particular, the MLF predicts that the determiner is sourced from the ML (i.e. the language of the finite verbs, which form the category of ‘Late System Morphemes’ as described in §4.2.1.1), while the MP predicts that determiners would be sourced from the language with grammatical gender (i.e. Welsh or Spanish, but not English). Their results show higher support for the MP in terms of coverage of the data, and no significant differences between the accuracy of the models’ predictions on both language pairs. However, they also argued that as Welsh and Spanish were also more frequently the ML (compared to English), and since Welsh in particular was the ML in all mixed DPs of this dataset, it would be inadequate to conclude that the preference for Welsh determiners was caused by the fact that Welsh has grammatical gender. By virtue of the language of the verb in the clause containing the mixed DP, they concluded that ‘the success of the Minimalist account was due to the fact that the language of the verb was almost always Welsh or Spanish, i.e. ‘a language with grammatical gender’ (p.571). The steps that were taken to reach this conclusion, however, remain unclear. Given that the language of the finite verb (i.e. the ML) also happens to be the language with grammatical gender, these two separate conditions have become structurally intertwined. In other words, if we refuse to believe that the preference for Welsh or Spanish determiners is a direct result of the grammatical gender feature encoded in said languages (as per the MP predictions), we similarly have no direct reason to believe that it was due to the ML (i.e. the language of the verb) either (as per the MLF predictions). In order to reach such a conclusion, what we need is a dataset where these conditions can be properly teased apart.

Similarly, adjective-noun combinations in several languages are another area where the predictions of the MLF and the MP contradict one another. Specifically, MacSwan’s MP builds on Cinque’s (1994) proposal of a Universal Base structure, in terms of which adjectives underlyingly start in a position *c*-commanding their nouns, which would lead to adjectives being spelled out in pre-nominal positions unless something else (e.g. movement) occurs. MP thus postulates that the conflict between the Universal Base and the noun-adjective surface order in some languages is a result of an overt movement of the noun to a position above the adjective. On this ground, MP predicts that the language of the adjective determines the word order of a mixed NP. Myers-Scotton’s MLF, in contrast, would predict that the language of the finite verb (i.e. which is

<sup>53</sup><http://bangortalk.org.uk/speakers.php?c=siarad>

<sup>54</sup><http://bangortalk.org.uk/speakers.php?c=miami>

associated with Late Outsider Morphemes in Myers-Scotton's terms) determines the word order. Several studies have found evidence for the MP (e.g. Cantone & MacSwan, 2009 for German-Italian, Wyngaerd, 2017 for French-Dutch), while others reported strong accuracy for the MLF predictions (e.g. Parafita Couto, Deuchar & Fusser, 2015 for English-Welsh). One of the latest studies in this area is that of Parafita Couto & Gullberg (2017), who directly test these two alternative theoretical accounts on 80 early English-Spanish bilinguals, recruited through Amazon Mechanical Turk. In these experiments, participants were asked to judge sentences under four different conditions: sentences that followed the CS pattern predicted by either the MP (but not the MLF), or the MLF (but not the MP), both, or neither. Results show that switches that conform to both (MP+/MLF+) are the most preferred, while switches that negate both (MP-/MLF-) are the least preferred. There were no significant differences between MP+/MLF- and MP-/MLF+ sentences. When participants were forced to choose which sentence was more 'acceptable,' results showed a preference for the MP predictions over the MLF, although MP-/MLF+ condition is favoured over the MP-/MLF- condition. It seems then that there might be some interactions between the predictions, but the evidence remains inconclusive. Additionally, whether or not these results are borne out in naturalistic corpora is also debatable. In fact, there has been contradicting evidence as to whether or not patterns found in experimental data are aligned with those in natural speech (e.g. Parafita Couto et al., 2015; Wyngaerd, 2017).

Ultimately, what we have seen thus far is that although the MLF has long been claimed as a 'universal model,' it is possible to find evidence both for and against its predictive power.

#### 4.3.1.2 The MLF in 'inconvenient' language pairs

It is also clear from the review that a common characteristic of all the studies so far is that they involve language pairs that have different word orders and/or different inflectional morphology (i.e. the two criteria upheld by the MLF principles). As far as I am aware, only one study to date has tested the MLF in a language pair that shares the same word order and has limited morphology. This is a study by Wang (2007), who tested the original MLF model on Mandarin-Tsou and Mandarin-Southern Min. While the MLF model could be straightforwardly applied to Mandarin-Tsou due to their different word orders and morphology, the model was found to struggle with Mandarin-Southern Min, a language pair that shares the same surface clausal word order with limited overt morphology. In particular, Wang (2007) found that the MLF principles were only able to account for less than 10% of the Mandarin-Southern Min data (N<30/300), with success rates ranging from 3-8% between groups of speakers. Wang overcame this problem by suggesting two additional criteria, the **Morpheme Counting Principle** and the **Uniform Structure Principle**, both first put forward by Myers-Scotton herself in earlier work (Myers-Scotton,

1993). Specifically, Wang suggested that if Myers-Scotton's original two principles—the System Morpheme Principle and the Morpheme Order Principle—do not work, one should resort to the Morpheme Counting Principle, which states that whichever language supplies the greater number of morphemes in the clause is considered the Matrix Language of that clause. If this principle is of no help either (e.g. in cases where the number of morphemes are approximately equal), the Uniform Structure Principle should be applied. The Uniform Structure Principle postulates that 'a given constituent type in any language has a uniform abstract structure and the requirements of well-formedness for this constituent type must be observed whenever the constituent appears' (Myers-Scotton & Jake, 2000a, p.120). In other words, it predicts that early and bridge system morphemes are supplied by the ML, because they favour keeping the structure uniform across the CP. Wang (2007, pp.214–215) used the following examples to illustrate this.<sup>55</sup>

- (25) a. *i er-zi a-bue bi-iap e kuan o* [Southern Min-Mandarin]  
 her son not-yet graduate NOM seem part-affirm  
 'Her son has not graduated yet.'
- b. *i giã a-bue bi-iap e kuan o* [Southern Min]  
 her son not-yet graduate NOM shape part-affirm  
 'Her son has not graduated yet.'
- c. *ta er-zi hai-mei bi-ye de yang-zi o* [Mandarin]  
 her son not-yet graduate NOM shape part-affirm  
 'Her son has not graduated yet.'

(Wang, 2007, p.214, *italics* = Mandarin, **boldface** = Southern Min))

In order to identify the ML of the bilingual clause in (25a), the two original principles of the MLF must first be checked. However, Wang points out that since the monolingual clauses in (25b) and (25c) have exactly the same word order, and no Late Outsider Morphemes are found,<sup>56</sup> neither the Morpheme Order Principle nor the System Morpheme Principle is applicable. In this case, Wang (2007) suggests that we look at the number of morphemes each language has supplied. As all except one word in (25a) comes from Southern Min, Southern Min is deemed the ML. Where the number of morphemes is almost equal as in (26), however, the ML is determined by the language of the Early System Morphemes (i.e. classifiers) and Bridge Morphemes (i.e. possessive markers).

<sup>55</sup>Glosses and translations are reproduced as per the original. As for why the glosses for *kuan* in (25a) and (25b) are different, Julio Song (a native speaker of Mandarin) helps to explain that *kuan* is the Southern Min counterpart of Mandarin *yang-zi*, both of which can mean either 'shape, appearance, look' (the literal meaning) or 'seem' (the metaphorical meaning).

<sup>56</sup>In his dissertation, Wang (2007) identified Late System Morphemes as subject-verb agreements or case affixes, the language of which determines the ML. However, since none of these features exists in monolingual Mandarin or Southern Min, it is not clear why such features should be expected in bilingual clauses combining these two varieties.

- (26) *i jing-zhui sen ɕʒi-liab zhong-lío.* [Southern Min-Mandarin]  
 His neck grow one-Sclass4 tumour  
 ‘There is one tumour on his neck.’

(Wang, 2007, p.215, *italics* = Mandarin, **boldface** = Southern Min))

Here, since the Early System Morpheme, the CLS *-liab*, comes from Southern Min, Southern Min is the ML. As the Vietnamese possessive marker and classifier function very similarly to those of Mandarin and Southern Min (§4.6.2), the Uniform Structure Principle could be applied to Vietnamese in the same way. However, as we shall see, neither the Uniform Structure Principle nor Morpheme Counting Principle is particularly helpful in the case of Vietnamese-English, for reasons I detail in §4.5.4.

In the case of Vietnamese-English CS, the application of the MLF has only come up very briefly in one study thus far (Tuc, 2003). Set in Victoria, Australia in 1994, the study posed a broad question of how first-generation Vietnamese speakers use CS in their bilingual repertoire. Tuc’s most relevant finding was that within a single NP, when a single attributive adjective is switched to English, the position of the switched English adjective is ‘always on the right-hand side of the Vietnamese noun’ (p.65), which reflects Vietnamese NP word order and contrasts with English. However, due to lack of information on how the ML was determined, this observation alone cannot show whether or not the structure of these mixed NPs conforms to the constraint prescribed by the MLF (i.e. the language of the finite verb dictates the word order of mixed Adj-N combinations). This finding also only constituted a small part of Tuc’s thesis, and therefore did not merit any further analysis in his study.

At this point, it becomes apparent that the MLF has been applied to a range of different language pairs, with various degrees of success. However, what seems missing in these studies is the due consideration of an important question that is not often asked: whose ML is it that we are talking about when we ‘assign’ the ML? Noticeably in these studies, the assumptions seem to be that community’s monolingual ‘norm’ is consistent with ‘standard’ facts of the participating languages, which are then unquestioningly used as a default point of reference. This is particularly problematic, given that it has long been established that language variation is the norm, not the exception (e.g. Labov, 1995; Tagliamonte, 2007; Mesthrie & Bhatt, 2008; Saraceni, 2010; Tagliamonte, 2011, 2012; Saraceni, 2015; Hudson Kam, 2015; Clark, 2016; de Vogelaer & Katerbow, 2017; Bolton, 2018; Poplack, 2018). Although Myers-Scotton addresses this by insisting that the ML is not ‘to be equated with an existing language’ but rather an ‘abstract construct’ for the morphosyntax of the bilingual CP (Myers-Scotton, 2002, p.66), this makes the operationalisation of the ML even more dubious. Specifically, she claims that the structural requirements imposed by

the ML need not be exactly the same as the source language, and this lack of ‘congruence’<sup>57</sup> can account for the occurrence of EL islands and EL bare forms (i.e. cases that are not optimally morphologically integrated into the ML) (Myers-Scotton, 2002, p.67). Unfortunately, there is little explanation of how ‘congruence’ between an ML and an EL can be defined, especially when the ML ‘does not include actual morphemes nor is it isomorphic with any fully fleshed-out linguistic variety’ (Myers-Scotton, 2002, p.68). As Auer & Muhamedova (2005) point out, the problem becomes obvious: without any point of reference, we simply have no means to establish morpheme order or late system morphology, which in turn renders the notion of an ML difficult to interpret.

### 4.3.2 Previous work on the Matrix Language Turnover Hypothesis

Having discussed previous work using the MLF, I will now review studies concerning its related model, the Matrix Language Turnover Hypothesis. As previously introduced, the ML Turnover Hypothesis was proposed as an explanatory mechanism for language shift, referring to situations where ‘the main language which had structured constituents becomes the structurally minor or Embedded Language (EL); in turn, the language which had been the minor language regarding structure becomes the ML’ (Myers-Scotton, 1998, p.300). Three ML Turnover scenarios are possible: (i) an arrested ML turnover, (ii) a Composite ML, and (iii) a complete turnover (§4.2.2).

It is worth noting, however, that while the MLF model has been relatively widely-applied, studies concerning the ML Turnover Hypothesis have been very limited by comparison. According to Myers-Scotton (2002), this might well be due to the fact that progressive grammars of earlier stages of languages left behind in shifts are ‘not typically available’ (p.248). This means that we cannot find any direct evidence that the route leading to shift was through an ML turnover. Though admitting that the ML Turnover Hypothesis has not been properly tested, Myers-Scotton insists that the hypothesis is eminently testable, provided that ‘longitudinal data of the relevant sort were collected’ (p.249).

The first study that was able to do so was Fuller’s (1996), using secondary data from previous research on Pennsylvania German (PG) conducted in the 1940s and in the late 1970s/1980s. Comparing the two datasets at different points in time, Fuller claims that an ML Turnover is underway, with a ‘Composite ML’ arising which carries features of the two languages in contact. Features of convergence are primarily found in the tense system, morphological realisation

<sup>57</sup>The idea of congruence is introduced in the 1993 version of the MLF, as part of ‘The Blocking Hypothesis.’ The Blocking Hypothesis states: ‘In ML+EL constituents, a blocking filter blocks any EL content morpheme which is not congruent with the ML with respect to three levels of abstraction (§4.2.2) regarding subcategorisation’ (Myers-Scotton, 1993, p.120, my own cross-reference). Accordingly, each EL element must be checked for ‘sufficient congruence’ between the ML and the EL before being inserted. In the absence of this, we see EL bare forms or EL islands (§4.2.1.2), which are a lawful violation of the MLF.

within VP, and ‘the increased syntactisation’ of word order. These ‘converged’ features are declared on the basis that they resemble more closely English features and somewhat depart from other German varieties. Fuller claims, for example, that because Plain PG uses [dative/locative preposition *am* + infinitive] while English uses [be + participle] to construct progressives, the construction in the following mirrors English sentences and shows abstract influence from English:

- (27) *Sie js am ready waerre fer ins Versammlung gen* [Plain PG]  
 she be.3so on ready become.iNF for in.NEU.ACC church go  
 ‘She’s getting ready to go to church.’

(Burrige 1992, p.213, cited in Fuller, 1996, p.503)

Accordingly, Fuller (1996) argues that the increased frequency of the progressive structure we see in (27), considered with the fact that ‘the SG [Standard German] equivalent would employ the simple present tense,’ indicates convergence towards English. Furthermore, as the use of this construction to express progressive aspect was also documented in Earlier Pennsylvania German (Frey, 1942; Buffington & Barba, 1954), Fuller concludes that we can assume that this convergence ‘is part of the first phase of the ML turnover’ (1996, p.503).

The solidity of such a conclusion, however, is questionable, not least because despite the fact that ‘frequency’ is repeatedly mentioned as evidence for structural convergence, no quantitative indication is given as to how frequent these constructions are. As researchers broadly agree, in order to declare convergence in a principled fashion, we need to be able to account for both cases where the structures ‘converge’ and where they do not (Labov, 1969, 1972)<sup>58</sup>, particularly when non-convergence cases are already deemed to exist in the literature (Fuller, 1996, p.503). Second, it seems unclear why Standard German is taken as a baseline for analysis, despite Fuller’s own admission that, ‘a study of language change must provide a comparison to illustrate that change has taken place, and today’s Standard German (SG) is not a valid basis of comparison.’ (Fuller, 1996, p.501).

A more recent study following Fuller’s vein is that of Kheir (2019), who compares different conversational datasets from the years 2000 and 2017 featuring Palestinian Arabic–Israeli Hebrew CS in the Druze community in Israel. Six out of 10 of the 2017 recordings included the same participants from the dataset collected in 2000, allowing the researcher to directly track change in individual speakers. In essence, the study makes two major claims:

<sup>58</sup>Labov’s Principle of Accountability (Labov, 1972, p.72) states that: ‘any variable form (a member of a set of alternative ways of “saying the same thing”) should be reported with the proportion of cases in which the form did occur in the relevant environment, compared to the total number of cases in which it might have occurred.’ Unless this principle is followed, it is possible to prove any theoretical preconception by citing isolated instances of what individuals have been heard saying. Speech is perceived categorically, and linguists who are searching for an invariant, homogeneous dialect will perceive even more categorically than most. The problem is most severe in the study of non-standard dialects.’ (Labov, 1969, p.737, n.20).

- (i) CS among the Israeli Druze has been changing over the years from ‘classic CS’ (CS with a dominant ML) to Composite CS; and
- (ii) this turnover has not gone to completion, but has created a split language along the way.

Kheir (2019) cites the following examples to support her case. Underlined are Hebrew-derived elements; others are Arabic, in **boldface** are the morphemes under discussion.

- (28) *Slixa inno tʔakhar-et heik pašūt kan fi ktir **pkak-āt** ʕ-ṭariq.* [Hebrew-Arabic]  
 Sorry that be-late-1SG-PST like-that simply was in a-lot traffic-PL on-the-way  
 ‘Sorry that I was late, there was simply a lot of traffic on the way.’

(2000 data, p.497)

- (29) *Mbareh roḥ-et ʕala **el-xanoot** **ve-štar-et** hai **el-simla** **lal-ʕores** **tabaʕ**  
 Yesterday go-1SG-PST to the-shop and-buy-1SG-PST this the-dress for-the-wedding of*

*ʕAnan.*

*ʕAnan*

‘Yesterday I went to the shop and bought this dress for Anan’s wedding.’

(2017 data, p.501)

According to Kheir (2019), in (28), a Hebrew masculine noun (*pkak*) is inflected with Arabic feminine plural suffix (*-āt*) and thus forms a ‘hybrid plural.’ She further noted that in Myers-Scotton’s terms, there is a distinction between ‘cultural borrowing’ and ‘core borrowing:’ the former covers lexical items that are new to the recipient language culture, whereas the latter refers to concepts that have viable equivalents in the recipient language (for examples of the distinction, see Nguyen, 2016, pp.14–15). Because there is ‘an equivalent’ in the recipient language, ‘core borrowing’ must be used for purposes other than filling a lexical gap, and hence only ‘core borrowing’ forms part of the structural borrowing identified in the ML Turnover Hypothesis (§4.2.2). Kheir thus analysed the word *pkak* ‘(traffic) jam’ in (28) as a case of a core borrowing, given Arabic has the viable equivalents *izdiḥam* ‘(traffic) jam’ and *izdiḥam-āt* ‘(traffic) jams’ (2019, p.497). In this sense, the Hebrew core borrowing has become lexicalised in the ML Arabic, by means of taking plural affixes according to the Arabic pattern. This was then taken as a sign of Phase I in an ML turnover.

Data from 2017, demonstrated in (29), however, shows a very different pattern: both languages seem to play a role in the syntax. As Kheir (2019) observes, the Hebrew content morpheme *ve* ‘and’ is often prefixed to Hebrew morphemes, but is now prefixed to an Arabic content morpheme *eštar-et* ‘bought’ while assimilating the *e* from both languages. Furthermore, the speaker formulates the Arabic possessive phrase according to the Hebrew pattern (‘for the wedding of Anan’ instead of the Arabic counterpart ‘for Anan’s wedding.’) On this basis, she diagnoses the occurrence of convergence.

This analysis is again problematic. First of all, the distinction between cultural borrowing and core borrowing is rather blurry, and having an ‘equivalent’ in the other language is a rather cryptic yardstick to assign the status of such elements. As I demonstrate in §4.5.4, whether an item could be considered a ‘core borrowing’ or a ‘cultural borrowing’ is determined both at a community and an individual level. For example, a lexical item might well have ‘an equivalent’ in the other language, but if the speaker had never acquired this ‘equivalent’ in that language, there is no ‘equivalent’ for them (Aikhenvald, 2002, p.197). Without information on the speakers’ acquisition background, and without reported variation in the corpus showing the frequency and diffusion of such items, we have no principled way to ascertain their status. In the case of *pkak* above, if it turns out not to be a ‘core borrowing’ per se once other conditions are taken into account, then it does not count towards the ‘code-switching-borrowing’ continuum, thereby not fitting in the description of Phase I in the ML Turnover Hypothesis.

Second, given the contact setting of Palestinian Arabic–Israeli Hebrew, both of which were previously described in the paper as a distinct language subgroup with influence from multiple ancient and modern languages (Kheir, 2019, pp.484–488), speakers’ monolingual codes should not be taken for granted. Although several ‘standard facts’ about the spoken varieties of these languages were given, it was not clear where these facts were drawn from, or whether they truly reflect the variety spoken by these speakers. Additionally, we do not have enough facts given such that we could conclude, for example, that the pattern in (29) is indeed a convergence with Hebrew rather than a result of internal variation or evolution of Arabic itself.

Another particularly relevant study that has also made use of the ML Turnover Hypothesis is that of Wang (2007) on Mandarin-Tsou.<sup>59</sup> Applying the MLF model to 130 bilingual clauses, Wang (2007) found that most of those collected from older Tsou people have Tsou as the ML (79%, N=79), while clauses produced by younger Tsou speakers predominantly have Mandarin as the ML (67%, N=20), i.e. indicating an ‘ML turnover’ in place. Wang (2007) further reports that while no example of innovation or borrowing is found in the data, there are instances of an omission of the tense/aspect marking in monolingual Tsou in the corpus. These omissions of tense/aspect marking, as shown in (30), are taken by Wang as evidence of structural borrowing from Mandarin.

- (30) a. *bonədo fou* (?o *Basuya*) [Monolingual Tsou recorded]  
 eat Obl2 meat Nom4  
 ‘(Basuya) ate meat.’

<sup>59</sup>This is the same study in which Wang studies Mandarin-Southern Min (§4.3.1.2). However, only Mandarin-Tsou is considered using the ML Turnover Hypothesis, which is why it is the sole focus of this section.

- b. *mo bonədo fou (?o Basuya)* [Grammatically correct Tsou]  
 tense2Agent eat Obl2 meat Nom4  
 ‘(Basuya) ate meat.’

(Wang, 2007, p.251)

Here, Wang suggests that because it is obligatory to mark tense in Tsou (30b) while it is not in Mandarin, the fact that the tense marker is omitted in (30a) is ‘evidence of structural influence’ from Mandarin as the majority language (Wang, 2007, p.252). On this basis, Wang argues that an ML Turnover has occurred. Given that most utterances produced by younger Tsou speakers are in monolingual Mandarin, he further suggests that the ML Turnover in the Tsou community has gone to completion. However, such a conclusion seems hasty, first of all because the frequency of these cases is not reported. Similar to the study by Fuller (1996), we do not know how often this occurs in the corpus, and how its distribution compares to cases where the tense markers are **not** omitted. Furthermore, the problematic nature of assuming speakers’ monolingual code is amplified by the facts that:

- (a) Tsou speakers were described as indigenous tribes living in mountainous areas, isolated from the mainland (Wang, 2007, p.18); and
- (b) the linguistic situation in Taiwan is rather complex, where ‘different varieties of Chinese are spoken, and the language of each Austronesian aboriginal tribe varies’ (Wang, 2007, p.22).

In summary, none of the studies using the ML Turnover Hypothesis thus far provides convincing enough evidence for the ML Turnover as a mechanism of change. Although the model is helpful in explaining variation patterns, the main issue in these studies is the prescriptively sanctioned standard monolingual baseline adopted as a point of reference (§4.3.1). Convergence or structural borrowing in these cases is simply evidenced by a deviation from ‘standard grammaticality’ of the languages involved, intuited by the researchers rather than based on the data produced by the speakers themselves (Torres Cacoullos & Travis, 2018). For data as variable as CS (Clyne, 1987; Gardner-Chloros & Edwards, 2004; Gardner-Chloros, 2009; Chan, 2009; Wang & Liu, 2013), this represents a significant gap in our understanding of the language contact as it happens. With data containing both speakers’ monolingual and CS utterances, CanVEC offers an opportunity to fill in this gap. In addition, no study to date has examined the ML Turnover Hypothesis in a modern migration setting, or on Vietnamese-English in contact.<sup>60</sup>

Before we examine whether an ML Turnover has occurred, however, we first need to establish whether it is possible to systematically establish the ML for each bilingual CP. The following section hence explains the protocol on the basis of which we proceed.

## 4.4 Establishing the Matrix Language in CanVEC

### 4.4.1 The System Morpheme Principle

As previously discussed, the System Morpheme Principle (§4.2) specifies that the language of the Late Outsider Morpheme is the ML of any given bilingual CP. The augmented 4M model (Figure 4.1) further stipulates that Late Outsider System Morphemes include any morpheme under INFL which cannot be realised without checking with another element in the sentence (i.e. ‘outside of Maximal Projection of Head.’) However, because Vietnamese does not inflect, this presents a significant challenge. In particular, Vietnamese has neither subject-verb agreement nor case affixes. Although Vietnamese makes use of the copula *là*, its form is also invariant. This is illustrated in examples (31)–(34) below.

- (31) Heather<sub>1</sub>: *cô-ấy là người được.*  
 3SG COP person ASP.Acquired  
 ‘She is the winner.’  
 (Heather.Troy.0807, 04:53.2–04:54.4)

- (32) Mina<sub>1</sub>: *nếu-như con là Bill Gates,*  
 If 2SG.kin COP Bill Gates  
 ‘If you were Bill Gates,’  
 (Mina.Pete.0905, 06:01.9–06:04.4)

- (33) Taz<sub>1</sub>: *em là người to blame.*  
 1SG.kin COP person  
 ‘I am the person to blame.’  
 (Tee.Taz.0808, 04:25.8–04:27.3)

- (34) Tim<sub>1</sub>: *nếu-như-mà họ là idol của con,*  
 if 3PL. COP of 2SG.kin  
 ‘If they were your idols,’  
 (Tim.Jess.Chloe.0705, 15:03.4–15:08.5)

As we can see, the form of the copula *là* is unchanged regardless of subject information (i.e. first-, second-, third-person singular in (31), (32), (33) respectively, and third-person plural in (34)). It is clear from these examples that there is no overt agreement between the subject and the

<sup>60</sup>It is worth pointing out that even though CanVEC data is not a traditional type of ‘longitudinal data’ as specified by Myers-Scotton (2002), it offers a proxy to this requirement. In particular, we have first-generation speakers representing an older variety and second-generation speakers representing a newer variety of Vietnamese as the heritage language. In fact, a similar approach was also demonstrated in Wang (2007), as we have previously seen. Empirically, this ‘longitudinal proxy’ using a cross-generational synchronic corpus is also in line with the general consensus that, in a migrant context, change can only be appropriately construed by comparing the speech of second- and third-generation immigrants with that of first-generation speakers (Poplack & Levey, 2010; Otheguy & Zentella, 2012).

copula in Vietnamese. This means that the System Morpheme Principle, as it stands, can only be applied to mixed clauses with an English finite verb.

Another issue with the System Morpheme Principle is that even when we have a mixed clause with an English finite verb, subject-verb agreement varies due to a phonological characteristic typical of Vietnamese L2 speakers of English. Specifically, the coda is most often unreleased, as in the following example:

- (35) Harry<sub>1</sub>: *mà* he live in America.  
           but  
           ‘But he lives in America.’

(Harry.Tressie.0508, 08:37.2–08:41.8)

Here, because the present-tense singular morpheme ‘-s’ following the verbal stem ‘live,’ i.e. the coda, was not realised, it was not possible to determine from the recording whether agreement had occurred or not. Similar examples of deleted codas in speakers’ production are well-represented across the corpus, and can be observed by listening to almost any recording. This observation is in line with previous studies’ discussion of Vietnamese speakers’ phonotactic tendency to delete or reduce final consonants in speech (e.g. Osburne, 1996; Lardiere, 1998; Patil, 2008), both in English and in Vietnamese. This phenomenon was also well-documented in one of the first education guides for Vietnamese refugees in the United States who were learning to speak English (National Indochinese Clearinghouse, 1977). An L2 phonological regularity therefore further reduces the applicability of the principle, not only to mixed clauses with an English finite verb, but also to those produced by second-generation speakers (where L2 effects do not play a role), or, in the case of those produced by first-generation speakers and where the subject is 3SG, only when an agreement is overtly realised (for 3SG). Where there is no phonetic realisation of a Late Outsider Morpheme when we expect one (as in (35)), we have no basis to determine whether the cause is down to phonology or syntax; only the latter, however, matters in establishing an ML.

At this point, one might argue that an area where the System Morpheme Principle manifests in a more clear-cut fashion is where the inflections are not regular and result in a change in vowel quality (e.g. ‘think’—‘thought’, ‘eat’—‘ate’ and so on). Because Vietnamese does not inflect at all while English does, whether the main verb changes in form (or not) can indicate which grammar constitutes the ML. Specifically, if irregular verbs change in form, it indicates English as the ML; if not, it points to Vietnamese as the ML. The issue with this diagnostic, however, is that it is also possible for speakers to use present tense to refer to past events in English, especially in lively narratives and quotations (Schiffrin, 1981; Singler, 2001; Tagliamonte, 2004, 2007). This so-called ‘historic present’ tense can be easily found in the corpus, as example (36) illustrates. Here, all the irregular verbs that result in vowel change in past tense are in **boldface**, and numbers in

the square brackets on a separate line indicate the number of intervening clauses not relevant to the point being made.

- (36) a. Reece<sub>1</sub>: but they **take** us in,  
 b. they cook Thai food for us.  
 c. and we at the cabin with the captain as well,  
 d. have a shower,  
 e. **sit** in there,  
 f. eating <E>,  
 g. **lie** down in the open.  
 h. anyway because we **speak** English,  
 i. and the captain of the Thai ship is very friendly,  
 j. [3]  
 k. he **sing** in Vietnamese.  
 l. [5]  
 m. they are transport ships,  
 n. they **come** to Vietnam sometimes,  
 o. so they pick up all the popular song,  
 p. the song everyone **know**.  
 q. oh they **sit** there sing,  
 r. ‘*Saigon đẹp lắm Saigon ơi Saigon ơi*.’  
 s. Taylor<sub>2</sub>: the song is about how beautiful Saigon is?

(Reece.Taylor.0906, 30:42.7–31:43.8)

As we can see, present-tense irregular verbs used to denote past events can be found in almost every line in this extract, from both first- (Reece, line a.–r.) and second- (Taylor, line s.) generation speakers. In fact, over half of the eligible irregular English finite verbs in CanVEC (i.e. as demonstrated by **boldface** verbs in example (36) above) were realised as an apparent present-tense form (55%, N=420/760). This casts English irregular inflection as a point of strong variation in the corpus, thereby rendering its diagnostic unreliable.

Poor display of overt inflections in both languages thus makes the System Morpheme Principle extremely problematic for a language pair like Vietnamese-English. The problem is further amplified by the fact that, similarly to other analytic languages (e.g. Mandarin, Cantonese, Thai), morphemes in Vietnamese are highly multi-functional; the same element, unchanged in form, can either be a system or a content morpheme depending on its distribution. Examples (37) and (38) illustrate:

- (37) Tanner<sub>1</sub>: *đi Hội-An ở resort* **đã** *lắm đó.*  
 go Hoi-An stay satisfying very DM  
 ‘Staying at resorts in Hoi-An is very satisfying.’

(Tanner.Nina.0609, 03:30.4–03:32.2)

- (38) Taz<sub>1</sub>: *thì hẳn đã làm* homework early.  
 then 3SG PST do  
 ‘Then he would have done the homework early.’

(Tee.Taz.0808, 02:48.7–02:50.6)

In (37), *đã* is a content morpheme that means ‘satisfying,’ which occurs after the clausal subject and acts as the main (stative) verb. In (38), however, the same element precedes the main verb *làm* ‘do’ and is a past tense marker.<sup>61</sup> Using the System Morpheme Principle in Vietnamese-English mixed speech, then, actually requires consideration of the morpheme position, rather than the form of the morpheme itself. This leads us to the discussion of the Morpheme Order Principle.

#### 4.4.2 The Morpheme Order Principle

The Morpheme Order Principle states that the surface word order of a mixed clause is determined by the ML. The application of this principle assumes that the two languages involved have different word orders at a clausal level, which is not the case for English and Vietnamese, however. Both languages are strictly SVO, meaning it is not possible to determine which language the word order comes from, as in (39) below.

- (39) Tyler<sub>2</sub>: *anh xem trailer,*  
1SG.kin watch  
 S V O  
 ‘I watched the trailer.’

(Billy.Tyler.Ellie.1108, 15:35.5–15:36.8)

The monolingual equivalents from Vietnamese and English are also provided in (40) for comparative purposes:

- (40) *anh xem đoạn-giới-thiệu,*  
1SG.kin watch trailer  
 S V O  
 ‘I watched the trailer.’  
 S V O

As (39)–(40) show, it is not possible to attribute the word order identified in (39) to either English or Vietnamese, as SVO is the default word order in both languages.

It is worth noting, however, that while the Morpheme Order Principle does not work at a clausal level, it is quite productive in the nominal domain and in the formation of interrogatives. Here I discuss how the Morpheme Order Principle could be applied to these cases in the corpus.

<sup>61</sup>Note that the English translation using future perfect in past tense ‘would have’ is based on information from the surrounding discourse. There is nothing in the syntax here to distinguish a simple past reading (‘did’) from the past future perfect (‘would have done’).



#### 4.4.2.2 Polar questions

The formulation of polar questions is another promising diagnostic that can be used to differentiate between English-ML and Vietnamese-ML clauses. Specifically, the inherent multi-functionality of Vietnamese means that the same forms are used in different syntactic positions to serve different purposes. Vietnamese polar questions constitute a structure in which this multi-functionality emerges: these questions feature the negator *không* or *chưa* at the end of the sentence. This is illustrated by the following example from the corpus:

- (44) Twee<sub>2</sub>: *me còn mượn mấy cái sách tiếng-Việt cho nó không?*  
 2SG.kin still borrow PL CLS book Vietnamese for 3SG NEG  
 ‘Do you still borrow Vietnamese books for him?’

(Theresa.Twee.0715, 14:46.3–14:55.5)

This structure differs from English, which involves moving an auxiliary verb (‘be,’ ‘do’ or ‘have’) or a modal into sentence-initial positions, followed by the subject (for example, see the English translation in (44)). The main difference between a Vietnamese polar question and an English polar question, then, lies in the position of the question-signalling word: i.e. movement of an auxiliary to the initial position of a clause in English, and the obligatory negation marker at the end of a clause in Vietnamese.

Consider (45)–(46) as examples:

- (45) Tressie<sub>2</sub>: *could ba rent it?*  
 2SG.kin  
 ‘Could you rent it?’

(Harry.Tressie.Josh.0719, 02:32.4–02:37.7)

- (46) Hannah<sub>2</sub>: *có nhiều assignments không?*  
 have many NEG  
 ‘Are there many assignments?’

(Hannah.Lida.0718, 10:02.8–10:03.9)

In (45), the modal ‘could’ was fronted to form the question and so it is clear that it follows English word order. English is then considered the ML. In contrast, the speaker in (46) inserted the negator *không* at the end of the utterance. This mirrors Vietnamese word order and so Vietnamese is considered the ML of the utterance.

#### 4.4.2.3 Wh-questions

The formation of Wh-questions is another syntactic feature that is not shared by English and Vietnamese. English content questions are typically formed by placing the Wh-word at the beginning of the utterance. In case the Wh-element is a non-subject, the Wh-phrase is followed by an auxiliary verb (Erickson, 2001). By contrast, Vietnamese is a Wh-in-situ language, which

means that Wh-questions follow the same SVO order as a declarative, with the Wh-element appearing in the position that would contain the answer (Nguyen, 1997). Examples (47) and (48) demonstrate these differences. In (48), an equivalent declarative is also provided in (b.) for clarification.

(47) Tim<sub>1</sub>: what is your ideal type?

(Tim.Jess.Chloe.0705, 00:54.3–00:56.7)

(48) a. Dany<sub>1</sub>: Ø *đang ở phòng của ai đó?*  
 PRG stay room POSS who DM  
 ‘Whose room (are you) at?’

(Brian.Dany.0812, 10:29.9–10:32.5)

b. Ø *đang ở phòng của tôi.*  
 PRG stay room POSS 1SG  
 ‘(I am) staying in my room.’

Such differences in the syntactic structures of Wh-questions suggest that in a mixed content question, the position of the Wh-word could reveal the ML. Consider the following example:

(49) Lida<sub>2</sub>: so what did you *học* in Health?  
 learn

‘So what did you learn in Health?’

(Hannah.Lida.0718, 00:24.0–00:27.1)

In this example, since the Wh-question was formed based on the English rule; i.e. the Wh-word placed at the beginning of the sentence and the movement of the auxiliary ‘did,’ English is tagged as the ML of this clause.

Having discussed the universal principles of the MLF and where they can be applied, I next present the outcome of applying these principles to CanVEC.

### 4.4.3 Results

It is clear from our previous discussion that there are only limited instances where the original MLF criteria can be applied to Vietnamese-English. Specifically, the System Morpheme Principle can only be applied to mixed clauses produced by second-generation speakers (due to first-generation speakers’ typical phonotactically driven deletion of coda ‘-s’), and the Morpheme Order Principle only works in the nominal and interrogative domain. Another point to make explicit here is that the System Morpheme Principle and the Morpheme Order Principle should be applied simultaneously. The outcomes of these two criteria are expected to converge, as example (50) illustrates.



‘no matter what languages are involved’ (Myers-Scotton, 2006, p.248). This is furthermore at odds with some previous suggestions that the MLF works unfailingly on genetically distinct languages (such as English and Vietnamese), where ‘there seems to be a universal tendency to select one morpho-syntactic rule’ (or, more precisely, the morpho-syntactic rules of one language) over the other (Chan, 2009, p.197).

As Vietnamese and English share clausal word order with limited or no inflectional morphology in play, an obvious-seeming explanation would be that most of these ‘difficult data’ follow from the relevant structures involving congruent word order. Figure 4.4, however, clearly shows that this is not the case.

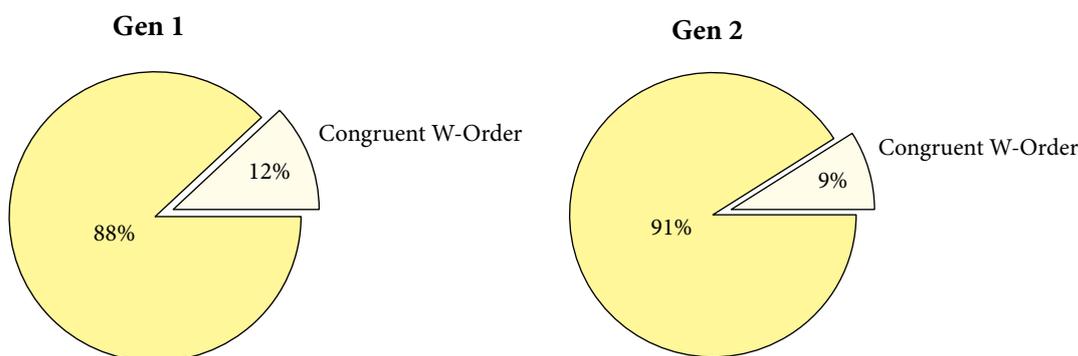


Figure 4.4: The proportion of ‘ML-ambiguous’ clauses with congruent word order

As we can see, only roughly 10% of the data without an identifiable ML involved congruent word order. A major proportion of this subset (~90%) remains unaccounted for. An appropriate question to ask at this point, then, is what else underlies this subset (N=1,555); or in other words, what makes it particularly challenging for the MLF to handle?

## 4.5 Difficult data

### 4.5.1 Whose Matrix Language is the Matrix Language?

Recall from §4.4.1 that we were only able to apply the System Morpheme Principle to mixed clauses with an overt realisation of English verbal agreement. The issue for the first-generation speakers, however, is that there is very little overt verbal agreement marking due to their phonotactic characteristics (§4.4.1). In cases where there is no overt agreement where it should occur, it is thus difficult to determine whether it is due to the speakers’ phonotactic tendency to delete final consonants, or whether it is because they actually do not have any syntactic agreement in place.



In the context of Reece's speech, one might argue that, though variable, the fact that he occasionally has agreement in his monolingual English may suggest that this feature does exist in his English variety. Given that Vietnamese does not require agreement at all while English does, overt agreement in his mixed clauses can thus only indicate that this grammatical feature comes from English. English should therefore be considered the ML. I would like to point out, however, that even if subject-verb agreement does exist in Reece's English, data clearly shows that his (and most CanVEC first-generation speakers') overt realisation and non-overt realisation of English agreements varies to a large extent, with the non-overt agreement being considerably more common in the corpus (N=212/247).<sup>62</sup> This observation suggests that there might be different factors conditioning overt and non-overt subject-verb agreement in speakers' English. Until we know precisely what these conditions are, however, we still do not have a strong basis upon which we can reliably assign English as the ML of the speaker's mixed utterance. For example, if the overt realisation of subject-verb agreement only occurs with verbs that end in a vowel such as 'go' in (52) in the speaker's English (which is particularly likely, given that Vietnamese does not allow consonant clusters, see e.g. Nguyen, 1997), then the fact that agreement occurs with a verb ending in a consonant such as 'sleeps' in (51) in the speaker's code-switching does not suggest that this late outsider morpheme comes from English. Instead, what it then indicates is a somewhat hybrid feature that might be considered a form of a 'composite' structure within the MLF (§4.2.2). This again amplifies the inherent problematic assumption of a prescriptively sanctioned standard monolingual baseline in interpreting the ML, a point I previously discussed in §4.3.1.

To sum up, the implication for first-generation speakers in CanVEC is thus that the System Morpheme Principle only successfully applies to the limited cases where the speaker consistently shows overt agreement in their production of English (N=2/28). For the majority like Reece, however, we do not have enough conclusive evidence to reliably assign English as the ML for these utterances (N=89), given the extent to which their English agreement varies.

## 4.5.2 Composite Matrix Language

Another issue that emerges from the difficult dataset is cases where the System Morpheme Principle and the Morpheme Order Principle each points to a different ML. The MLF states that in any given mixed CP, only one language is the ML, and only this ML can supply the word order and Late Outsider Morphemes for the utterance. On this basis, the System Morpheme Principle and the Morpheme Order Principle are non-hierarchical criteria, and their results should ultimately converge. However, it is apparent from the data that this is not always the case. For example,

<sup>62</sup>Remember that the raw counts are not extremely high because they are limited only to first-generation speakers' monolingual English, and only to third-person present tense here.

there is some evidence of word order coming from one language but Late Outsider Morphemes from the other. Consider the following:

- (54) Hannah<sub>2</sub>: the *thầy* just makes us like go through *cái* booklet *này*.  
 teacher CLS DEM  
 ‘The teacher just makes us like go through this booklet.’  
 (Hannah.Lida.0718, 01:04.0–01:11.3)
- (55) Hannah<sub>2</sub>: and it was *cái* mistake worst of my life.  
 CLS  
 ‘And it was the worst mistake of my life.’  
 (Hannah.Lida.0718, 15:45.9–15:53.9)<sup>63</sup>
- (56) Hannah<sub>2</sub>: and it was *cái* competition really *bự này của* ABC Triple Six *đó*.  
 CLS big DEM POSS DM  
 ‘And it was this really big competition of ABC Triple Six.’  
 (Hannah.Lida.0718, 26:03.9–26:08.1)
- (57) Twee<sub>2</sub>: because the thing *khó* is,  
 difficult  
 ‘Because the difficult thing is,’  
 (Theresa.Twee.0715, 08:05.7–08:06.9)

As examples (54)–(57) illustrate, there is a pattern of conflicting evidence for the ML, when the Late Outsider Morpheme comes from one language but the morpheme order comes from the other. Specifically, the finite verbs in these examples all come from English, and the overall clausal word order could be English (or Vietnamese; we cannot tell due to congruent word order), yet the nominal word order follows that of Vietnamese: ‘booklet *này*’ (N + DET) in (54), ‘mistake worst’ (N + ADJ) in (55), ‘competition really *bự này*’ (N + ADJ + DET) in (56), and ‘thing *khó*’ (N + ADJ) in (57). Given that both Vietnamese and English contribute to the basic grammatical structure, we are left with no means to determine the ML of these utterances.

Recall from §4.2.2, however, that this ‘Composite ML’ phenomenon is expected as one of the possible outcomes of an ML turnover, in which the CP shows ‘mixed’ grammatical features in one way or another. Accordingly, it is necessary not only that a Composite ML exists, but also that it has ‘fossilised into the main medium of communication.’ However, a potential Composite ML only makes up a small proportion of the difficult data in CanVEC. In fact, out of 1,555 clauses without an identifiable ML, only four were found to be Composite (<1%). This means that the second scenario within the ML Turnover Hypothesis does not apply here: it would require a major proportion of the clauses to exhibit composite structure and become the major medium of communication in the community.

<sup>63</sup>Note that this example can be interpreted as two separate clauses with ellipsis, i.e. ‘And it was the mistake, (which is) worst of my life.’ However, as there is no prosodic break between ‘mistake’ and ‘worst,’ we can safely conclude that this is indeed one coherent clausal unit, and should treat it accordingly.

Another crucial aspect that is worth pointing out is that, though small in numbers (N=4), all of the Composite-ML clauses were produced by second-generation speakers. Given that Myers-Scotton (1998, p.299) attributes the mechanism of a Composite ML to second language learners (whose ‘imperfect knowledge’ of the target language licenses ‘Composite-ML structures’), we would expect that most of the Composite-ML utterances were produced by the first-generation speakers who acquired English as a second language. This is not the case in CanVEC. Intriguingly, while this result contradicts Myers-Scotton’s (1998) theoretical proposition, it is consistent with her own empirical findings in a study of the Arab-American community in Columbia (Jake & Myers-Scotton, 2009). In particular, she showed that second-generation Arabs produced eight times more Composite-ML clauses than first-generation speakers (N=34 vs. N=4). The theoretical implication of these findings, however, was not explicated, leaving the conceptual-empirical disconnect that bedevils the notion of ‘Composite ML’ unresolved.

Ultimately, the point worth stressing here is that the current MLF model cannot account for cases where the results of the Morpheme Order Principle and the System Morpheme Principle are in conflict. While the idea of a Composite ML has been offered as a way out, the concept is inherently dubious (§4.2.2). Although it seemingly addresses cases that show the lexicons from one language but the grammar from another, the concept of a ‘Composite ML’ still cannot offer a principled explanation for cases exhibiting mismatched grammatical outcomes (e.g. (54)–(57)). Simply labelling these instances Composite-ML structures therefore seems like stretching the model to cover data that does not fit.

### 4.5.3 Clauses with null elements

The majority of the ‘difficult data,’ in fact, is characterised by language-specific constructions that neither of the MLF principles is capable of accounting for, in particular clauses with null elements (N=1008). The first to be discussed are dropped English auxiliaries (e.g. ‘Ø you finished yet?’, or ‘What Ø you looking for?’; see Caines & Buttery, 2010; Caines et al., 2016). Despite the repeated claims that the MLF model is designed to accommodate spoken data (Carter, Deuchar, Davies & Parafita Couto, 2011; Deuchar et al., 2018), it does not prove to be particularly effective on CanVEC production.<sup>64</sup> Specifically, CanVEC speakers drop auxiliaries approximately

<sup>64</sup>While writing has been traditionally taken as the primary source of grammatical description, speech has a tendency for simplified and disjunctive construction, with grammatical structure playing a lesser role in the overall communication process (Leech, 2000; Leech & Svartvik, 2003; Buttery, McCarthy & Carter, 2015; Ginzburg & Poesio, 2016; Carter & McCarthy, 2017). Numerous studies have characterised features that are predictive of the spoken grammar repertoire, with a particular focus on phenomena that are markedly more frequently or differently distributed in spoken discourse (see Cheshire & Fox, 2016 for work on prefabricated expressions and affective meanings, Fried & Östman, 2005; Crible & Cuenca, 2017 on discourse markers; Weinert & Miller, 1996; Wagner, 2010; Cresti, 2014; Spronck & Nikitina, 2019 on coordination and subordination in speech; Mair, 2013; Čermáková, Komrsková, Kopřivová & Poukarová, 2017 and Zhang, Li & Luo, 2018 for cleft constructions and conjunctions).

one fifth of the time in monolingual English (~20%, N=51/288).<sup>65</sup> Examples (58)–(60) below demonstrate this variation. Consistent with the convention elsewhere, occurrences of auxiliaries are underlined, while absences are marked by inserted Ø.

- (58) a. Reece<sub>1</sub>: because of course they will ask you,  
 b. where do you want to go?  
 c. how long Ø you going,  
 d. and when Ø you coming back?

(Reece.Taylor.0906, 03:53.3–04:00.7)

- (59) a. Nina<sub>2</sub>: the first thing she told me was,  
 b. we Ø going to Da Nang,  
 c. and the second thing she told me like two weeks later,  
 d. oh we Ø going to Saigon as well.  
 e. and after that like two weeks later she said,  
 f. oh and we are staying in Hoi An as well.

(Tanner.Nina.0609, 03:57.8–04:09.5)

- (60) a. Jess<sub>2</sub>: you Ø giving me a different standard,  
 b. because I am a woman.  
 c. Chloe<sub>1</sub>: he Ø not giving standard.  
 d. (6)  
 e. Jess<sub>2</sub>: obviously I am going to.  
 f. I am going to live my life with my own rules,  
 g. whether I am a man or a woman.  
 h. I am just saying,  
 i. let me try it once twice.  
 k. Tim<sub>1</sub>: no I Ø not want you to try it.  
 l. Jess<sub>2</sub>: well you need to let me try it.  
 m. Tim<sub>1</sub>: I do not think <X>,  
 n. Jess<sub>2</sub>: dad I know.  
 o. It is not good,  
 p. but I Ø probably going to try it at some point.

(Tim.Jess.Chloe.0705, 12:32.3–13:33.2)

As we can see, auxiliary drop occurs quite frequently in spoken English across different generations. This phenomenon is also prevalent in CanVEC mixed clauses, as shown in examples (61)–(63).

<sup>65</sup>This proportion is calculated against the total and is limited to only the envelope of variation, i.e. clauses where an auxiliary is expected to occur, rather than the grand total of all clauses in monolingual English. Cases of ‘subject-auxiliary’ deletions such as ‘been there’, ‘done that’ were also excluded, as they have become so idiomatic that we cannot assume that they vary in the same way as sole auxiliary drop.

- (61) a. Harry<sub>1</sub>: yeah he Ø getting a bit from *bố Phát với-lại mấy* uncle *đó*.  
 father Phat together-with PL DM  
 ‘Yeah he (was) getting a bit from father Phát and the uncles.’
- b. Tressie<sub>2</sub>: they Ø all coming from *Huế bố Phát anh Đào?*  
 Hue father Phat brother Dao  
 ‘They (were) all coming from Huế, father Phát (and) brother Đào?’  
 (Harry.Tressie.Josh.0719, 22:11.0–22:22.0)
- (62) a. Taz<sub>1</sub>: *ví-dụ-như bây-giờ em* Ø trying my best,  
 for-example now 1SG.kin  
 ‘For example (if) I (was) trying my best now,’
- b. *em* Ø helping *nó*,  
 1SG.kin 3SG  
 ‘I (am) helping him,’
- c. *em* Ø given *nó mười cái* tools.  
 1SG.kin 3SG ten CLS  
 ‘I (have) given him ten tools.’
- d. (2)
- e. *chớ-không-phải là em* Ø offering *mười cái* tools,  
 not COMP 1SG.kin ten CLS  
 ‘Not that I (am) offering ten tools,’
- f. *là em* expect *là*,  
 COMP 1SG COMP  
 ‘Then I expect that,’
- g. *hắn phải* use *được hết mười cái* tools.  
 3SG. must acquire all ten CLS  
 ‘He must (be able to) use all ten tools.’
- h. *nhưng-mà em nghĩ*,  
 but 1SG.kin think  
 ‘But I think,’
- i. at least *là em* Ø given *nó được mười cái* tool.  
 COMP 1SG.kin 3SG acquire ten CLS  
 ‘At least I (have) given him ten tools.’  
 (Tee.Taz.0808, 07:32.1–07:51.8)
- (63) a. Henry<sub>2</sub>: *con bọ gì* Ø bitten *phát thành* vegan *luôn ấy*.  
 CLS bug what shot become DM DM  
 ‘There are some bugs that (have) bitten (you) once (and then you) become a vegan.’
- b. *thì nó sẽ có một loại gọi là* virus,  
 then 3SG will have one type call COP  
 ‘Then there will be a type called virus,’
- c. *nó* Ø stuck ở *trong máu của chú*,  
 3SG LOC inside blood POSS 2SG.kin  
 ‘It (is) stuck inside your blood.’  
 (Tom.Henry.0809, 48:52.0–49:08.0)





Traditionally researchers often determine such shared grammatical features by referring to the monolingual norms of the participating languages (see §4.3.2). For example, in the Pennsylvania German case, changes in word order (closer to that of English and further away from other German varieties) were taken as evidence for ‘convergence’ or a composite structure at work (Fuller, 1996). By that logic, cases of null elements here would fall into the ‘Composite’ category, given that they are much more widely permitted in Vietnamese than in English. In §4.3.2, however, I observed that this approach is problematic for its treatment of speakers’ varieties as homogeneous. In fact, as we will see in Chapter 5, the predisposition for null arguments in particular is more complex than it seems in Vietnamese, and different speakers are under different constraints as to when they can and cannot leave elements unexpressed. It thus seems unsatisfactory to classify the clauses under discussion here as either English-ML or Composite-ML clauses, based on the current principles. The ultimate point of a problematic prescriptively sanctioned standard monolingual baseline adopted by the MLF is hereby again made, this time illustrated by data from CanVEC.

#### 4.5.4 A note on Wang’s additional principles

In the last section of my discussion on difficult data, I would like to return to Wang’s proposal to extend the MLF model to challenging language pairs (2007; 2016). Specifically, Wang previously reported the limited applications of the MLF on isolating languages (Mandarin-Southern Min), and proposed the re-introduction of the ‘Uniform Structure Principle and the ‘Morpheme Counting Principle’ as a solution (§4.3.1). In this section, I will discuss how using these principles for CanVEC is still problematic.

First, consider the Uniform Structure Principle, which states that Early and Bridge Late System Morphemes come from the ML as the unmarked choice—‘just because it gives preference to keeping structure uniform across the CP’ (Myers-Scotton, 2002, p.120). In this sense, the bridge morpheme *của* (possessive marker) in Vietnamese could potentially be used to test this principle; however, it is optional in most contexts. Example (68) demonstrates a typical case in the corpus, where the Bridge Morpheme *của* is not phonologically realised at all.

- (68) Sony<sub>1</sub>: *trời mấy trường Ø bạn Ø em ở Việt-Nam* [...(a long VP)]  
 God PL school (POSS) friend (POSS) 1SG.kin LOC Vietnam  
 ‘God, all my friends’ schools in Vietnam [had a six- or seven-AM start].  
 (Quentin.Sony.0306, 016:06.0–16:09.3)

In fact, this observed phenomenon reflects the widely permitted omission of the Bridge Morpheme in Vietnamese (Nguyen, 1997, p.184): *của* is only required where the possession is to be emphasised or contrasted. Line (f.) in example (69) illustrates:

- (69) a. Dany<sub>1</sub>: *bên đây học khác ở Việt-Nam.*  
 side DEM study different LOC Vietnam  
 ‘Studying here is different to in Vietnam.’
- b. Brian<sub>1</sub>: *mhm dạ đúng.*  
 DM right  
 ‘Mhm that’s right.’
- c. *bên này không-phải là lên mạng rồi tìm,*  
 side DEM NEG COMP on Internet then search  
 ‘It is not about looking and searching on the Internet.’
- d. *rồi search,*  
 then  
 ‘Then search.’
- e. *rồi copy and paste,*  
 then  
 ‘Then copy and paste.’
- f. *tất-cả là ý-kiến của mình.*  
 all COP idea POSS 1PL.REFL  
 ‘Everything is our own idea.’

(Brian.Dany.0812, 09:12.4–09:21.5)

Furthermore, even when Bridge Morphemes do occur, they pattern similarly to the Vietnamese Early System Morpheme *cái* (Vietnamese generic classifier) in that they are present mostly in monolingual Vietnamese clauses (69f), English clauses with Late Outsider Morphemes already present (70a), or mixed clauses that are already ML-identifiable by the original principles (70b).

- (70) a. Hannah<sub>2</sub>: *so focus của my group is,*  
 POSS  
 ‘So the focus of my group is,’
- b. *if the movie is better than cái cuốn book.*  
 CLS-GENERAL CLS-SPECIFIC  
 ‘If the movie is better than the book.’

(Hannah.Lida.0718, 06:53.4–07:02.4)

We also do not have any case where the English Bridge Morphemes (e.g. ‘of’) or the Early System Morphemes (e.g. articles ‘a,’ ‘an,’ ‘the’) are found where no other principles could reveal the ML. The Uniform Structure Principle is thus found to be inapplicable to CanVEC.

We next consider the Morpheme Counting Principle, which posits that: ‘The ML is the language of more morphemes in interaction types including intra-sentential CS’ (Myers-Scotton, 1993, p.68). As Myers-Scotton argues, ‘the language that is the source of the grammatical frame often supplies more morphemes in a bilingual CP’ (2002, p.61). In other words, the ML determined by the Morpheme Counting Principle should be consistent with that identified by the Morpheme Order Principle and the System Morpheme Principle. When neither the System Morpheme Principle nor the Morpheme Order Principle applies, the Morpheme Counting Principle

has been suggested as a gateway to identifying the ML (Wang, 2007). It is, however, important to note that while the Morpheme Counting Principle enabled Wang to resolve ‘most of the Mandarin-Southern Min bilingual clauses’ (Wang, 2007, pp.210–211, numerical rate not available), it presents several problems that are difficult to ignore in the present study.

The first issue in applying the Morpheme Counting Principle is that frequency counts cannot be applied to individual clauses. As Myers-Scotton observes, ‘frequency counts must be based on a discourse sample; they offer no reliable evidence if they are performed on single sentences’ (1993, p.68). Wang’s decision to apply them at the level of individual clauses for Mandarin-Southern Min is therefore rather puzzling. Furthermore, it is worth noting that the concept of a ‘discourse sample,’ in fact, has not been well-defined either. It remains unclear what counts as a discourse sample, and ‘how large is large enough is an unresolved issue’ (Myers-Scotton, 1993, p.68). On the whole, another crucial empirical question to ask is whether the language that has the larger total number of morphemes in the entire corpus should be considered the ML, irrespective of the intricacies of word order and system morphemes at a clausal level. If so, this seems to defeat the idea of the ML as a ‘grammatical construct.’ If not, we need an explicit explanation of how and under what constraints the principle can operate.

Second, Myers-Scotton also states that ‘cultural borrowings from the Embedded Language for new objects and concepts are excluded from the count’ (Myers-Scotton, 1993, p.68). Yet the question of what counts as borrowing is already controversial in the CS literature, let alone the distinction between ‘cultural borrowing’ and ‘core borrowing.’<sup>66</sup> According to Myers-Scotton (1993, pp.168–171), because there is ‘an equivalent’ in the recipient language, ‘core borrowing’ must be used for purposes other than filling a lexical gap (§4.3.2). Only in this case is borrowing considered a valid morpheme for the Morpheme Counting Principle. On an empirical basis, this distinction appears unclear as there is no specification of how, or in what aspect, a concept can be considered to have a ‘viable equivalent’ in the other language. I have previously reported this difficulty in a separate study of the usage of Vietnamese kin terms in the Vietnamese-Australian bilingual community (Nguyen, 2016), using the following example:

- (71) *sao* you *đặt* Ø *Huy* với *Duy*?  
 why name Huy with Duy  
 ‘Why did you name (them) Huy and Duy?’

(Transcript H, 11:54.9–11:57.1)

Here, I argued that determining whether something is a ‘cultural borrowing’ or a ‘core borrowing’ is less than straightforward. Semantically, ‘you’ has an equivalent in Vietnamese (*mày*), yet these can hardly be considered pragmatically or culturally equivalent. While ‘you’ is a neu-

<sup>66</sup>Note that this problem had not been dealt with at the stage of language marking. Recall from our discussion in Chapter 3 that CanVEC only marks established borrowing based on frequency and diffusion. It does not mark which borrowing was ‘cultural,’ or which was ‘core,’ as per Myers-Scotton’s distinction.

tral pronoun in English by which the speaker's relationship with the interlocutor is not specified, *mày* in Vietnamese is an inappropriate pronoun to refer to senior interlocutors. This argument also applies to many Vietnamese kin terms, which have semantic but not pragmatic equivalence in English. The question, then, remains whether 'you' in (71) should be classified as a core borrowing or a cultural borrowing for the purpose of an ML analysis.<sup>67</sup>

Finally, as Muysken (2000) points out, the Morpheme Counting Principle is questionable when applied to language pairs that are typologically disparate with respect to morphology. On Myers-Scotton's Swahili-English CS data, he writes: 'an agglutinating language like Swahili encodes many grammatical concepts (which are crucial structurally) with an overt morpheme, while isolating languages often do not' (2000, p.66). In other words, applying the Morpheme Counting Principle on a language pair involving typologically distinct morphological systems is problematic because it will favour the one with the larger inventory of grammatical morphemes. This point was well-taken by Myers-Scotton (2002), and contributed to her decision to abandon this criterion in later work. Although this was not a problem for Wang (2007, 2016) as he argued that Mandarin and Southern Min are both isolating, that is not the case for Vietnamese-English. Specifically, although both Vietnamese and English are morphologically limited (Chapter 1), English still has moderate inflection marking person-number, tense and aspect, whereas Vietnamese has no obligatory grammatical device for doing so. As we can see in example (72), a Vietnamese exact equivalent (Luna, 72c) of a simple English clause (Tressie, 72a) has fewer overt morphemes due to lack of verbal agreement ('-s' in the English verb 'likes'). Similarly, Vietnamese often also does not overtly mark tense (e.g. 72d & 72e) and hence the range of morphemes in the verbal domain is far more limited.

- (72) a. Tressie<sub>2</sub>: Alana likes you.  
 b. Luna<sub>1</sub>: yeah,  
 c. *Alana thích me.*  
    Alana like 1SG.kin  
    'Alana likes me.'  
 d. *xong-rôi trước-khi nó đi ngủ,*  
    then before 3SG go sleep  
    'Then before she went to bed,'

<sup>67</sup>It is also worth recognising that the distinction between 'core borrowing' and 'cultural borrowing' is determined both at a community and an individual level. In this study, where a speaker's language competence is self-assessed, it is difficult to know if a lexical item is borrowed to fill a legitimate 'lexical gap' in one language or another. For example, a speaker with lower English proficiency might borrow a word in Vietnamese to fill their own 'lexical gap' in English, despite there being an English 'viable equivalent' for that word. The criteria of 'having no viable equivalent' and 'being used to fill a lexical gap' are thus potentially in conflict, leaving us with no independent evidence either way. Therefore, the blurry line between cultural borrowing versus core borrowing makes it even more difficult to operationalise the Morpheme Counting Principle.

- e.           Ø *kéo cái* skirt.  
              pull CLS  
              ‘(She) pulled the skirt.’

(Luna.Tressie.0901, 07:55.7–08:04.1)

This, taken together with the fact that Vietnamese broadly allows null arguments (as in 72e; see also §4.5.3), means that applying the Morpheme Counting Principle to a language pair like Vietnamese-English is particularly difficult.

#### 4.5.5 Summary

In summary, data presented in this section has laid bare three essential problems with the MLF:

- (i) the problematic nature of taking speakers’ assumed monolingual code as a baseline;
- (ii) the lack of a principled strategy for cases where outcomes of the Morpheme Order Principle and the System Morpheme Principle are in conflict; and
- (iii) its struggle to deal with certain types of modern production data, especially those involving contemporary spoken colloquial features (e.g. zero auxiliary) and non-standard L2 features (which is a given in many bilingual contexts).

While these issues are not confined to just Vietnamese-English (see §4.3.1 and §4.3.2), the Can-VEC data has shown, on specific empirical grounds, the extent to which the model’s assumptions can be both quantitatively and qualitatively problematic. Despite Wang’s (2007) pioneering attempt to rescue the model for contact situations involving isolating languages, the proposed solution is not particularly helpful in the case of Vietnamese-English, given the added complexity of the fine-grained differences in the morphological typology of the languages involved. To summarise, Table 4.2 provides an overview of the distribution of data with which the MLF particularly struggles.

As Table 4.2 shows, across both generations, the majority of difficult cases involve null elements, including null finite verbs and English-ML clauses with Vietnamese abstract influence, i.e. Vietnamese-like null arguments. While it is not possible to proceed much further with null finite verbs (as we have no basis to determine which language the verb is ‘coming from’), null arguments are a topic that I will return to in Chapter 5. For now, having considered the nature of the large proportion of the data without a readily identifiable ML (57%, N=1,555/2,721), the crucial next step is to revert our attention back to the data where the ML was identifiable.

Type	Gen 1		Gen 2	
	N	%	N	%
Congruent word order without any other clue	108	12%	125	19%
No overt (agreement) late outsider morpheme without any other clue	216	24%	5	1%
Overt late outsider morphemes in mixed clauses of speakers whose English agreements vary	89	10%	0	0%
Contradicting outcomes of SMP and MOP	0	0%	4	1%
<b>Null elements</b>				
Null finite verbs	179	20%	210	32%
English ML with Vietnamese-like null arguments	307	34%	312	47%
<b>TOTAL</b>	<b>899</b>	<b>100%</b>	<b>656</b>	<b>100%</b>

Table 4.2: An overview of CanVEC difficult data in relation to the MLF

## 4.6 Matrix Language Turnover in the community

As explained in §4.2.2, an ML Turnover is a process whereby the original Matrix Language (i.e. the language which supplies the basic grammatical structure) of a bilingual CP becomes the Embedded Language and vice versa (Myers-Scotton, 1998). In most cases, the original ML is the minority language and the Embedded Language is the majority language. Here, I present CanVEC results as to whether such a turnover exists (§4.6.1), whether the phenomenon could be accounted for by the ML Turnover Hypothesis' predictions (§4.6.2), and highlight an alternative perspective for the data that seemingly does not fit in with the ML Turnover Hypothesis (§4.6.3). Readers should remember that we are only working with a limited number of clauses here, where the ML and EL could be firmly established (§4.4.3).

### 4.6.1 Is there a Matrix Language Turnover?

Following the definition of an ML turnover, the expectation is that in case an ML Turnover is observed, we would see a reversed distribution of the ML across the first and second generations, similar to what Wang (2007) found for Mandarin-Tsou. On this basis, Figure 4.5 reports the proportion of Vietnamese and English MLs in ML-identifiable mixed clauses across both groups of speakers.

As Figure 4.5 illustrates, it looks as if an ML Turnover has occurred. While Vietnamese is overwhelmingly the ML in most mixed clauses for the first-generation (78%), English has replaced this role for the second generation (53%). In other words, the old ML (Vietnamese) has lost its dominance as the ML, being replaced by the new ML (English) in second-generation

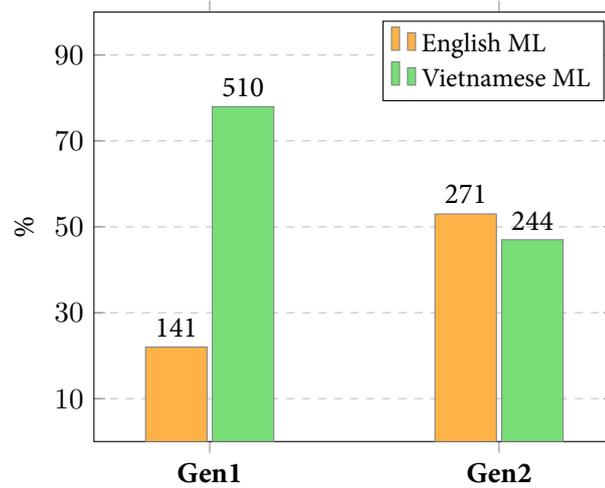


Figure 4.5: The opposite distribution of the ML across generations (CanVEC bilingual CPs)

speakers. Table 4.3 also further shows that the proportion of both monolingual English and English-ML mixed clauses is much higher in the second generation than it is for the first generation (30.7% vs. 18.8% and 6.9% vs. 1.9% respectively, contrast highlighted).<sup>68</sup> This cross-generational difference is statistically significant ( $\chi^2 = 120$ ,  $p < 0.01$ ), validating the hypothesis of a cross-generational turnover. This move towards the majority language is consistent with previous findings from other non-MLF-based studies on second-generation immigrants in Australia (e.g. Clyne, 2003; Karidakis & Arunachalam, 2016) and elsewhere (e.g. Ishizawa, 2004; Garcia-Colon, 2004; King & Fogle, 2006).

CP Type	Gen 1		Gen 2	
	N	%	N	%
Vietnamese monolingual CP	5,301	72.3%	2,207	56.2%
English monolingual CP	1,375	<b>18.8%</b>	1,207	<b>30.7%</b>
Mixed CP with Vietnamese ML	510	7.0%	244	6.2%
Mixed CP with English ML	141	<b>1.9%</b>	271	<b>6.9%</b>
CP with a Composite ML	0	0.0%	4	0.1%
<b>TOTAL CPs</b>	<b>7,327</b>	<b>100%</b>	<b>3,929</b>	<b>100%</b>

Table 4.3: Contrastive distribution of ML-identifiable CP types across generations, barring CPs whose ML cannot be identified (11%, N=1,555) and non-clausal IUs (N=8.8%, 1,236).

It is worth noting, however, that while the shift towards English among the second generation is not particularly surprising, what emerges as interesting is that this shift does not fit into any

<sup>68</sup>Note, though, that often only data from mixed clauses is taken as direct evidence for or against a turnover (e.g. Myers-Scotton, 2002; Wang, 2007).

of the scenarios as proposed by the ML Turnover Hypothesis. Recall from §4.2.2 that an ML Turnover can manifest in three different scenarios (Myers-Scotton, 1998, p.301):

1. The original ML is still maintained, but with some degree of structural borrowing from the other language.
2. A dual, Composite ML fossilises into the main medium of communication.
3. A complete turnover, where ‘CS falls away’ and production is characterised by ‘a single, standardised variety of what was the “new” ML during CS.’

Scenario (1) of an ML Turnover is the first to be ruled out, since Vietnamese (the old ML) is no longer ‘dominant’ and has been replaced by English (the new ML) for second-generation speakers. Scenario (2), the ‘Composite ML’ fossilisation, similarly does not fit, as there only seem to be four Composite-ML clauses in the dataset. This limited number of Composite-ML clauses makes up less than 0.1% of CanVEC production, and hence rather transparently cannot be described as ‘the main medium’ of communication in the community (see also §4.5.2). Lastly, the third scenario, ‘a complete ML turnover,’ is likewise ill-suited. As Table 4.3 shows, monolingual Vietnamese is still strongly present across generations, with first-generation speakers producing over 70% of their CPs in Vietnamese (N=5,301), and second-generation speakers producing 56% of the equivalent (N=2,207).

At this point, it is clear that the ML Turnover Hypothesis can neither predict nor account for what we see from the data in CanVEC. However, given that evidence for a ‘ML turnover’ is still quantitatively present, it is worth considering a further aspect of this model, i.e. the unidirectional effects from the ‘new ML’ on the ‘old ML.’

#### 4.6.2 Direction of structural borrowing

When a turnover is believed to have occurred, the ML Turnover Hypothesis predicts that any structural borrowing will manifest in the direction of the majority language. Recall from §4.5.3 that the augmented ‘Abstract Level Model’ posits three levels of abstract grammatical structure:

- (i) the lexical-conceptual level (semantic/pragmatic features);
- (ii) the predicate-argument level (relations between thematic role assigners—verbs and some prepositions—and the arguments they map onto phrase-structure units); and
- (iii) the morphological realisation level (elements and constituent orders surfacedly realised).

Given that there is evidence for an ML Turnover towards English, we would expect that any abstract change observed would be moving away from Vietnamese towards English. For example, we might expect to see Vietnamese clauses in English word order (where differences exist, §4.4.2), with subject-verb agreements, or even featuring functional elements such as definite articles and the like.

Qualitative analysis of the Vietnamese monolingual sentences in the corpus, however, shows no such abstract influence from English. On the contrary, it is striking that change is detected in the opposite direction: while novel elements such as articles or articles expressing definiteness were nowhere to be found in monolingual Vietnamese clauses, a handful of otherwise-English clauses were found to contain the Vietnamese generic classifier *cái*. Recall from Chapter 3, Table 3.2 that *cái* was used so frequently and widely as a single-word insertion that it constitutes one of the few borderline cases for established borrowing (frequency count = 9, cut-off = 10). Table 4.4 shows all of these instances. In this table, text inside [] is the immediately surrounding clause given for context.

Line	Speaker	Gen.	Transcript	Timestamp	Clause
a.	Reece	1	Reece.Taylor.0906	40:43.0–40:49.6	if you know someone working at <i>cái</i> butcher's shop,
b.	Harry	1	Harry.Tressie.Josh.0719	18:58.0–19:08.0	and dinner time or lunch time <i>cái</i> parents just open the door, [‘ <i>về ăn cơm mấy đứa.</i> ’]
c.	Hannah	2	Hannah.Lida.0718	06:53.4–07:02.4	[so the focus of my group is,] the movie is better than <i>cái cuốn</i> book.
d.	Hannah	2	Hannah.Lida.0718	08:23.7–08:27.5	<i>cái</i> movie is better.
e.	Hannah	2	Hannah.Lida.0718	26:21.3–26:26.6	and so we got to go on an <i>cái</i> excursion.
f.	Tressie	2	Harry.Tressie.Josh.0719	15:56.8–15:58.6	<i>cái</i> name <i>Dũng</i> is ugly.
g.	Taylor	2	Reece.Taylor.0906	38:32.2–38:35.5	so when you came to <i>cái</i> refugee camp,
h.	Twee	2	Theresa.Twee.0715	02:49.9–02:52.9	We were watching <i>cái</i> show Ninja Warrior.
i.	Twee	2	Theresa.Twee.0715	20:44.5–20:48.4	[behind there is a little play area,] like <i>cái</i> Belconnen but it is a lot bigger.

Table 4.4: All cases of the Vietnamese generic classifier in CanVEC otherwise-English clauses

In this instance, the ML Turnover model offers little explanatory power to account for the data under consideration here. Specifically, despite a possible ML Turnover being detected in the direction of English, not only is structural influence from English difficult to find, the strong presence of Vietnamese classifiers in these otherwise-English sentences suggests a somewhat opposite trend. While some might argue that classifiers are actually a subset of nouns carrying substantial semantic values (and are therefore content words rather than function words) (Nguyen, 1957; Cao, 2003), this analysis does not hold for several reasons. First, Vietnamese classifiers cannot occur on their own. Example (73) below, adapted from Tran (2011), illustrates this restriction.

Even though classifiers are similar to general nouns in that they can directly follow all adverbs indicating quantity such as *nhiều* (many), *ít* (little), numerals, quantifiers (*những*, *các*, *mấy*, which are all plural markers),<sup>69</sup> *vài* (several), *mỗi* (each)), indefinite article (*một*), demonstrative determiners (*này*, *kia*, *nọ*, *đó*) and Wh-word (*gi*), they cannot stand independently without other elements (73a). They must co-occur with at least a numeral (73b) or a demonstrative determiner (73d).

- (73) a. \**quyển*  
CLS
- b. *Tôi lấy hai quyển*  
1SG take two CLS  
'I'll take two.'
- c. *Tôi lấy quyển đó*  
1SG take CLS DEM  
'I'll take that one.'
- d. *Quyển đó dày lắm*  
CLS DEM thick very  
'That one is very thick.'
- e. \**Quyển dày lắm*  
CLS thick very
- f. \**Tôi lấy quyển*  
1SG take CLS

(Adapted from Tran, 2011, pp. 13-14).

Second, Vietnamese classifiers cannot function as independent subjects (73e) or direct objects (73f). Only when used as an anaphor (examples (73c) and (73d)), can classifiers take on grammatical roles as arguments. This is to say that classifiers differ from content nouns, and function akin to a grammatical category in many aspects. In fact, as Wang (2007) previously reasoned, the choice of a classifier is determined by its noun (shape, size, animacy, etc) and should thus be treated as an 'Early System Morpheme' according to the 4M model.<sup>70</sup> The presence of Vietnamese classifiers in otherwise-English clauses therefore indicates a separate mechanism that cannot be

<sup>69</sup>For further information on how these plural markers differ, see Nguyen (1997).

<sup>70</sup>Recall from Section 4.2.1.1 that Early System Morphemes 'depend on their heads for information about their forms...and are indirectly elected by their head content morphemes' (Myers-Scotton, 2002, p.75). In this sense, as classifier meets both of the requirements for being elected by its head noun, and its form is determined by the semantic content of the head.

grouped with the ‘lexical borrowing’ class.<sup>71</sup> As Myers-Scotton also insists, unlike content morphemes, system morphemes are not easily ‘borrowed,’ and ‘when we see system morphemes in a language, they are not the result of the same mechanisms that result in lexical borrowed forms, just with more cultural contact added’ (2002, p.244). She further attributes the presence of such system morphemes to ‘remnants’ of an arrested ML turnover.

It is important to note, however, that while we accept that CLS is a system morpheme and its presence in monolingual English indicates a different mechanism, the resulting explanation of an ML Turnover model still fails to account for what is happening here. Specifically, applying the ML Turnover model to the CanVEC data would predict that:

- (i) Vietnamese was taking over from English as the ML in the community;
- (ii) this turnover was arrested very early on; and
- (iii) Vietnamese classifiers are the ‘remnants’ of the proposed turnover.

We have seen, however, that (i) is simply not true (Table 4.3), which then in turn renders (ii) and (iii) untenable.

Having established that the MLF Turnover Hypothesis holds little explanatory power for the data at hand, the next question becomes how else can we best account for the occurrence of Vietnamese classifiers in otherwise-English discourse? I will next address this question from an acquisition perspective.

### 4.6.3 Early syntactic knowledge: A case of stability

In Table 4.4, what stands out is that **all** single word insertions of classifiers seen in the corpus are the general classifier *cái*. Given that Vietnamese has been said to have more than 200 classifiers (Nguyen, 1957; Truong, 2003; Cao, 2003), this raises an immediate question of what makes *cái* the prevalent choice.

In a comprehensive study on acquisition of Vietnamese classifiers, Tran (2011) found that the category CLS is acquired very early on, and children can accurately produce obligatory classifiers as early as age 1;11 across all combinations (CLS + N, CLS + DEM, CLS + Wh). This is consistent with cross-linguistic tendencies (cf. Hu, 1993 on Mandarin; Carpenter, 1987 on Thai; and Wong,

<sup>71</sup>Myers-Scotton (2002, p.242) also suggests, though, that in some cases, Early System Morphemes can be ‘borrowed’ along with their content morpheme heads. She cites Haugen’s (1950, p.218) example of English plural ‘-s’ being borrowed into American Norwegian ‘with its stem and treated as if it were part of a singular noun.’ In these cases, Early System Morphemes can be treated as part of a ‘loan word,’ thereby underlining a lexical borrowing process. This, however, is not the case here. As Vietnamese classifiers are inserted independently of their head nouns (which is always in English, Table 4.4), these classifiers are not ‘borrowed’ with their morpheme heads. They thus clearly differ from what is seen in the American Norwegian case.



tune with what was found in early stages of acquisition of French gender (Grégoire, 1971), when children first acquire unit-like pairings such as article-noun and erroneously add an additional article, understanding the actual article (pronounced with liaison) as part of the noun (e.g. *le loiseau* ‘the bird’).

It is important to note, however, that this explanation only holds in a CS context where the speaker systematically acquired ‘*cuốn* book’ as a fixed unit from her parents. This appears not to be the case, however, given that *cái* is typically the selected classifier to combine with English nouns in the corpus, as we have seen. The fact that we have double classifiers still might thus be accounted for by Tran’s (2011) findings that Vietnamese specific classifiers are acquired rather slowly for monolingual children, and even when a specific target classifier has entered a child’s repertoire, the use of the general classifier is not immediately abandoned. This means that infrequent, extended use of the generic CLS may still occur, even after speakers are deemed to have acquired the right classifier (i.e. *cuốn* in this case). This over-generalisation of the generic CLS is also borne out in (74c), where the general CLS *cái* is used together with the English head noun ‘excursion,’ which, in Vietnamese, is a type of noun that requires its own specific CLS (*cuộc*) and does not collocate with the generic *cái*.

Throughout all these ‘over-generalisation’ cases, what stands out is that speakers demonstrate stable syntactic knowledge of the CLS construction consistently. The phenomenon in (74c), for example, i.e. where *cái* follows the English indefinite article ‘an’ to form a complex NP (‘an *cái* excursion’), resembles the ‘adjacency principle’ in the numeral CLS construction in Vietnamese expressing indefiniteness. Specifically, in such constructions, the indefinite article (i.e. *một*, as the word for Vietnamese numeral ‘one’) together with a CLS forms an uninterrupted sequence with components that cannot be separated (Tran, 2011). Example (75) below illustrates this adjacency principle. As we see in (75a) and (75c), NUM-CLS remains together even when the associated NP is fronted and separated from the numeral and the classifier. Separating this sequence by an Adj (75b), or by moving the CLS together with the NP (75d) renders the sentence ungrammatical. For ease of interpretation, all instances of NUM and CLS here are in **boldface**.

- (75) a. Phoebe<sub>1</sub>: *bây-giờ mình vẫn chưa có **một cái** quyết-định cụ-thể*  
 now 1SG still NEG have NUM<sub>ONE</sub> CLS decision concrete  
 ‘Up until now I still have not made a concrete decision.’  
 (Mia.Phoebe.0905, 09:51.3–09:53.0)
- b. \**bây-giờ mình vẫn chưa có **một** cụ-thể **cái** quyết-định*  
 now 1SG still NEG have NUM<sub>ONE</sub> concrete CLS decision
- c. *quyết-định cụ-thể, bây-giờ mình vẫn chưa có **một cái***  
 decision concrete now 1SG still NEG have NUM<sub>ONE</sub> CLS  
 ‘A concrete decision, I still have not made one.’

- d.            *\*cái quyết-định cụ-thể, bây-giờ mình vẫn chưa có một*  
 CLS decision concrete now 1SG still NEG have NUM<sub>ONE</sub>

In other words, a Vietnamese classifier is obligatory when indefinite articles are involved. This constraint is strongly reflected even when the speaker produces otherwise monolingual English (as previously shown in example (74c)).

A crucial point worth noting here is that, while the tendency to over-generalise is not a new phenomenon (see e.g. Marcus, Pinker, Ullman, Hollander, Rosen, Xu & Clahsen, 1992; Schönenberger, 2001; Pérez-Leroux, Munn, Schmitt & DeIrish, 2004; Schuler, Yang & Newport, 2016; Yang & Montrul, 2017), what remains interesting is the subsequent retraction process. As Cournane (2019, p.143) suggests, ‘children generalise when they discover the basis for a rule or other systematic relation, and then gradually retract by learning sub-regularities, exceptions, blocking factors, or other factors governing the selection of one form (or meaning) over another.’ Although this over-generalisation process is believed to retract gradually as further, more complex input becomes available later on in the acquisition process (Biberauer, 2017, 2019; Cournane, 2019), the retraction slope varies, or retraction might not happen at all in some cases. For example, while Swiss-German children are found to retract their over-generalised embedded Wh-V2 by the age of five (Schönenberger, 2001), we have seen in (74a), (74b), and (74c) that over-generalisation of Vietnamese classifiers can survive in contact situations and remain into adulthood. This arrested retraction has also been similarly observed elsewhere, as in the case of the now-established feature of embedded-V2 in Wh-clauses in Afrikaans (Biberauer, 2019), or of learners’ extension of the definite plural marker *-ye* to non-specific nouns in Neo-Louisiana Creole (Mayeux, 2019).<sup>73</sup> This is not to say that the few instances of the over-generalised Vietnamese classifiers reported here could necessarily instantiate change scenarios like those mentioned above. What seems clear, however, is that over-generalisation can, under the right circumstances, produce diachronic change in contact situations. There is therefore the potential for the over-generalised Vietnamese classifiers reported here to ultimately feed into a long-term change. I leave closer investigation of this pattern to future research.

Overall, this section has shown that early syntactic knowledge of particular syntactic properties (here: of the classifier requirement, and, specifically, of the generic and animate instantiations of the classifier) probably plays a vital role in shaping what remains stable and what not

<sup>73</sup>In this study, Mayeux (2019) found that in Neo-Louisiana Creole (NLC), learners over-generalise the definite plural marker *-ye* to non-specific nouns, consistently using *-ye* as a general plural marker regardless of the specificity of the noun it modifies. Mayeux further reports a statistically significant difference between his 2012 and 2015 samples, and takes this as an indication for the preference of *-ye* having emerged over time. Given the lack of any inter-generational Louisiana Creole transmission in the home, Mayeux sees NLC (i.e. the variety of Louisiana Creole where the over-generalisation of *-ye* exists) as ‘maybe the sole incarnation of LC to be maintained over the next few decades’ (2019, p.102). This corroborates Biberauer’s and Cournane’s point about over-generalisation potentially leading to diachronic change, under the right set of circumstances.

over time. Unfortunately, this aspect is still inherently lacking in the MLF and the ML Turnover Hypothesis. Despite previous attempts to use the 4M model to classify the ‘natural order’ of acquisition (e.g. Wei, 2000; Namba, 2004),<sup>74</sup> the MLF itself and the ML Turnover Hypothesis have yet to incorporate the implications of such factors into its proposed mechanism of language change. As we have consistently seen throughout the chapter, any strong reliance on the reference grammar of a language without taking into account the nuanced intricacies of acquisition, the community, and the specificity of the varieties involved, runs the very real risk of operating with simplistic assumptions.

#### 4.6.4 A note on ‘stable bilingualism’

Before concluding the chapter, it is appropriate to deal with Myers-Scotton’s claim in her latest version of the MLF (2002) that the MLF is devised to account for ‘stable bilingualism’ only, and that ‘problematic cases’ often do not fall inside this remit (p.111). Specifically, she writes: ‘the MLF cannot account for all the structures in the CS of speakers in those communities where the relative status of the languages—in terms of both speaker proficiency and socio-political prestige—is more fluid than not,’ and then goes on to name ‘recent immigrants’ as a case in point (Myers-Scotton, 2002, p.111). Given that the Canberra Vietnamese bilingual community is a modern immigrant community, one might suspect that the MLF is not applicable to this contact setting to begin with, thereby explaining the lack of success of the model on CanVEC data.

It should be noted, however, although the modern migration situation is often seen as fluid, the situation in Canberra is quite different. First, the community is relatively small in size (1.6% of the total Vietnamese population in Australia) and attracts the lowest number of recent Vietnamese migrants due to its lack of a defined Vietnamese neighbourhood, high living cost, and more limited job opportunities in comparison to other diasporas (Australian Bureau of Statistics, 2017). This has produced a community comprising mostly young, educated speakers who acquired English and Vietnamese at a young age, either simultaneously or subsequently. Second, the fact that all speakers of CanVEC have been in Australia for at least 10 years (cf. Jake & Myers-Scotton, 2009) also means that speakers’ proficiency in both languages is likely to have surpassed

<sup>74</sup>Wei (2000) previously used the 4M model to formulate what he terms the ‘Hierarchy Principle’ in language acquisition, which stipulates that ‘directly elected morphemes (content morphemes) are acquired before system morphemes, and indirectly elected morphemes (Early System Morphemes) are acquired before structurally assigned ones (Late System Morphemes)’. He then conducted experiments in both Chinese and Japanese adult second language learners and found supporting evidence for this. Namba (2004) later also proposed the 4M model as an explanation for the ‘mysterious’ case of Brown’s acquisition order of three kinds of ‘-s’ in English: plural ending ‘-s’ > possessive ‘-s’ > third-person ‘-s’ (Brown, 1973). Namba argues that if we apply the 4M model and the Hierarchy Principle here, it is obvious that these three morphemes are acquired in the order of Early > Bridge > Outsider. However, while these attempts are helpful in classifying the order of morpheme acquisition, Namba (2004) did not make clear what values the model offers in explaining the mechanism of change.

the ‘unstable’ learning phase, or in other words, to have to some extent ‘fossilised’ into a stable ‘endstate grammar’ (Hawkins, 2000; Birdsong, 2004; Long, 2008). Several acquisition works have addressed this stabilisation in some detail, including longitudinal studies showing virtually ‘no changes’ in speakers’ grammars between the recording sessions conducted nine years or more apart (e.g. Lardiere, 2007; Long, 2008). Finally, as we have seen in [Chapter 2](#), the community is characterised by a dense social network, regular internal contact, and a high degree of communally shared information. All of these factors have been previously recognised as identifiers of a ‘stable bilingual community’ (Trudgill, 2011), thereby rendering ‘unstable bilingualism’ a weak explanation for the lack of success of the MLF model on the CanVEC dataset.

On a broader scale, it should also be noted that this caveat of ‘stable bilingualism’ for the MLF is also out of sync with the ML Turnover Hypothesis. Specifically, if the MLF only applies to stable bilingualism, how do we falsify the ML Turnover Hypothesis, which posits ‘dramatic shake-up in socio-political situations’ as a pre-requisite for any language change to occur? In other words, the conditions for different parts of the Matrix Language model do not match, leaving us with little room to account for various types of data and contact situations.

## 4.7 Chapter summary

This chapter has set out to probe cross-generational language variation and shift within the Canberra Vietnamese community, using the ML Turnover Hypothesis. Specifically, it probed Vietnamese heritage language indirectly by investigating its participation in the community bilingual discourse. In so doing, it tested how readily applicable the putatively universal MLF model is to Vietnamese-English in CanVEC, a new dataset that involves languages with homologous word order and extremely limited morphology. Results show that the ML Turnover Hypothesis and its associated MLF model only shed limited light on the ongoing changes within the community’s heritage language. Specifically, the MLF model fails to account for the majority of the CanVEC bilingual data, including both first- and second-generation speakers’ production (58% and 56% respectively). The CanVEC data also highlights the problematic nature of assuming speakers’ monolingual code as a basis of comparison, the ‘Composite ML’ notion, and the assumption of null elements in mixed discourse.

In relation to cross-generational ML turnover, results further demonstrated that even when the ML is putatively identifiable and evidence for an ML Turnover is quantitatively present in the community, we still do not find the kind of structural borrowing that the ML Turnover Hypothesis predicts. In fact, while English has seemingly taken over from Vietnamese as the dominant ML in the second generation, abstract structural influence is detected in the opposite direction.

Specifically, a significant proportion of the data exhibits Vietnamese-like patterns of null elements, and a handful of otherwise-English clauses is also found with the Vietnamese generic classifier *cái*. Of course, it remains a possibility that if all the data could be accounted for by the MLF, there might not be any turnover at all, i.e. that Vietnamese remains the dominant ML in the second generation; and if so, the fact that we see some abstract influence from Vietnamese on English may be explained via the ML Turnover Hypothesis. In any case, however, the conclusion remains transparent: the definitions of the MLF component parts are insufficiently clear, and, even if one tries to sensibly flesh out these components, the predictions do not seem to reflect what we see in CanVEC. The next natural step is therefore to ask, how else can we meaningfully probe the cross-generational language variation in heritage Vietnamese, without having to appeal to a Matrix Language? The pursuit of this goal is the focus of the next chapter.



# CHARACTERISING GENERATIONAL DIFFERENCES: A VARIATIONIST STUDY

---

## 5.1 Introduction

The attempt to probe Vietnamese heritage language indirectly via its participation in bilingual discourse in [Chapter 4](#) only gave a limited insight into cross-generational variation in the heritage language. In this chapter, I continue the enquiry by moving away from the MLF and the bilingual subset of the corpus to examine the Vietnamese heritage language monolingual subset directly. As null elements emerged as a distinct area of difficulty in [Chapter 4](#), I take the distinction between the null and overt realisation of functional elements as the focus of further investigation in this chapter. Specifically, I compare cross-generational patterns of three cases where null and overt alternation exists in Vietnamese: subjects, objects, and copulas. Given that English as the majority language features a system where overt forms are more strictly required for these three variables, the contact-induced change hypothesis would predict that changes are expected for subjects, objects, and copulas in the heritage variety.

The framework adopted in this chapter is the variationist approach (Labov, 1972 et seq.), which does not assume a ‘Matrix Language’ per se, but takes as central the regularity that underlies the variation of the languages as they are spoken within the community (Labov, 1972). The key advantage of the variationist approach is that it allows the heritage language to be examined as it is spoken in the community, without reference to any idealised benchmark. This not only holds significant descriptive value, but also allows us to identify trends and the direction in which the heritage language appears to be evolving. Crucially, the variationist focus on ‘community’s natural speech’ is coherent with the type of data that CanVEC comprises ([Chapter 3](#)).

This chapter consists of eight main parts. [Section 5.2](#) outlines the key principles of the variationist approach, [Section 5.3](#) provides the necessary background for subjects, objects, and copulas, and [Section 5.4](#) discusses previous work on these three phenomena in a cross-generational context. [Section 5.5](#) next lays out the coding method, while [Section 5.6](#) presents the results. [Section 5.7](#) discusses the implications of these results, before [Section 5.8](#) concludes the chapter.

## 5.2 Key principles of the variationist approach

In this section, I discuss two key elements of the variationist framework that bear direct relevance to the present study: the notion of orderly heterogeneity ([§5.2.1](#)), and the methodological innovations that have become a trademark of variationist sociolinguistics ([§5.2.2](#)).

### 5.2.1 Orderly heterogeneity

The variationist tradition takes as central the inherent variability of language use, where a ‘linguistic variable’ is a heuristic theoretical way of representing variability (Poplack, 1980; Milroy & Wei, 1995; Poplack & Meechan, 1998; Kiesling, 2005; Poplack & Levey, 2010; Tagliamonte, 2012; Tagliamonte & Baayen, 2012; Eckert, 2012). More specifically, linguistic variables are defined as structural instances where there are two or more ways to say ‘the same thing,’ where ‘the same thing’ refers to what is denoted by a form/an utterance (Kiesling, 2011, p.13). As an example, polar questions in English can be marked by either the typical subject-auxiliary inversion (e.g. ‘Do you like it?’), or by using declarative clauses with a rising intonation (e.g. ‘You like it?’). These two different constructions make up two variants for the variable of polar question marking in English.<sup>75</sup>

This emphasis on speakers’ choices foregrounds what variationist researchers know as **orderly heterogeneity**. In brief, orderly heterogeneity refers to the fact that although speakers of a language have different ways of saying the same thing (hence the heterogeneity), these choices are socially and linguistically structured (hence the orderliness). Consider Labov’s (1972) hallmark study of rhoticity in New York City as an example. There, each individual was found to behave differently in terms of post-vocalic /r/ production, yet which speakers and which utterances are more rhotic was predictable. In particular, the lower middle class (i.e. a social predictor) was seen to lead the spread of the prestige *r*-ful form, and this form is most strongly conditioned

---

<sup>75</sup>The notion of ‘the same thing’ can, in fact, be as challenging to operationalise as the concept of ‘a lexical gap’ ([Chapter 4](#), [§4.5.4](#)). However, the essential idea here is that there is ‘one isolable linguistic feature that carries meaning,’ and the community has more than one way of representing it (Kiesling, 2011, p.10). In this sense, the emphasis is thus on function and interpretive matters rather than form, or in other words, the choices that are available to speakers to achieve the same communicative intent.

by whether /r/ is followed by a consonant in the syllable (i.e. a linguistic predictor). Social predictors typically include broadly defined categories such as age, sex, class, education, and so on, while linguistic predictors vary according to the variables of interest. Individual speakers differing along social categories within a speech community are often expected to differ in their speech patterns. In the context of the present study, generational membership is taken as an important social predictor, and speakers belonging to different generations are expected to differ in their patterns as to when or how they choose a null form over an overt one.

### 5.2.1.1 Orderly heterogeneity in a focused community

It should be noted, however, that communities vary in different ways, and as such we cannot expect ‘orderly heterogeneity’ to be unanimously applied to different contact situations. As early as Le Page & Tabouret-Keller (1985) and Le Page (1989), a distinction was made between a ‘**focused community**’ and a ‘**diffuse community**’. A community is considered ‘focused’ when there is a high level of agreement on the shared speech norms within the community (hence more orderly heterogeneity), and ‘diffused’ when the speech norms are much less unanimous (hence less orderly heterogeneity). In other words, the extent to which a speaker receives feedback from the social environment concerning their language use determines the extent to which they can control and modify their speech in order to fit into that community. Strictly speaking then, the concept of ‘orderly heterogeneity’ should only be applied to communities that are highly focused.

Although the precise criteria for a ‘high level of agreement’ can be challenging to pin down, there are several indicators as to why the Canberra Vietnamese community should be considered focused. First, there is pressure to speak ‘good English’ (i.e. standard Australian English), particularly among first-generation speakers. For the second-generation, there is also pressure to speak ‘good Vietnamese’ (Chapter 2, §2.4.4). Standard Australian English and fluent Vietnamese are therefore relatively focused varieties that form part of speakers’ desired speech norms. Second, despite different political backgrounds, first-generation speakers often abstain from using words associated with *Tiếng Việt Cộng Sản* (the Communist Vietnamese variety) to avoid triggering any political tension (Chapter 2, §2.2). Finally, although speakers’ different backgrounds lead to split opinions on language attitudes towards each language, community members are quite undivided in their chosen identity: 90% (N=40/45) identify themselves as both Vietnamese and Australian rather than one or another (Chapter 2, §2.4.4). These facts, coupled with the well-established strong network of mutual support (Chapter 2, §2.3.2), demonstrates a high level of language and more general cultural agreement in the community, thereby justifying the decision to categorise CanVEC speakers as ‘highly focused.’

### 5.2.1.2 Orderly heterogeneity and individual agency

One of the main criticisms of the notion of orderly heterogeneity, however, has been its lack of consideration for the roles of individual speakers participating in the speech community. Dynamic concepts such as gender fluidities (Eckert & McConnell-Ginet, 1992), ethnic crossing (Rampton, 1995; Cutler, 1999) or styles (Bell, 1984; Coupland, 2007, 2009, 2011) continuously emerge as direct challenges to the Labovian predefined macro-sociological categories. Eckert (2012), for example, points out that categorising speakers on the basis of bundles of demographic characteristics is rather simplistic, neglecting the fact that speakers also have agency over the meanings which they want to create. Various studies have shown how particular forms can be used to achieve specific communicative and social functions, such as to disparage (e.g. Wong, 2005), to identify with a perceivably admired quality (e.g. Bucholtz, 1999), or to create an in-group membership (e.g. Zhang, 2005). This squares with Le Page & Tabouret-Keller's position that the individual 'creates for himself the patterns of his linguistic behaviour so as to resemble those of the group or groups with which from time to time he wishes to be identified, or so as to be unlike those from whom he wishes to be distinguished' (1985, p.181). In this view, speakers' linguistic variation does not just **reflect** their orderly position in a systematic structure, but also actively **constructs** it.

Eckert's (2012) critique is an important point to note as it brings to the forefront the role of individual styles, even in communities with a high level of linguistic cohesion. Various studies have shown how speakers draw from their available linguistic repertoire to extend, adapt, or invert social meaning of the form used (Bell, 1999; Schilling-Estes, 2004; Zhang, 2005; Coupland, 2007; Podesva, 2011). For example, Becker (2014), in her re-visitation of Labov's (1966) work on New York City English, found that not only is there a marked withdrawal from the New York City variants found to be dominant in Labov's study some 50 years back, but also many speakers are inconsistent in the usage of certain variables. As Becker (2014) observes, the use of more traditional variants is often present in situations when speakers want to identify themselves as local Lower-East siders, as opposed to the influx of non-local residents as the neighbourhood gentrifies. In these cases, the choice of a variant is not necessarily part of a systematic social differentiation, but rather a temporary and contextual choice.

This identity construction process has been referred to as bricolage (Eckert, 2004, 2008), a practice in which 'people combine a range of existing resources to construct new meanings or new twists on old meanings' (Eckert, 2004, p.42). The priority of individual agency in the bricolage model may seem to be at odds with an assumed coherent variety in the orderly heterogeneity model; however, these two approaches are not necessarily in conflict. In fact, as Eckert (2004, p.43) herself notes, since style is put out into a community to be interpreted, 'speakers select

resources on the basis of their potential comprehensibility in that community.’ This means that in order to successfully communicate any new meaning as part of the bricolage, speakers rely on shared understanding with their interlocutors on these existing resources and their potential indexicalities. Such shared understanding lies at the core of the orderly heterogeneity that Weinreich, Labov & Herzog (1968) advocate. Furthermore, even though the interpretation of a linguistic choice might be locally established, some level of consistency is required in order for the choice to be considered a style in its own right (Auer, 2002).

In the context of this work, the implication is precisely that while the notion of orderly heterogeneity is fundamental, it cannot be taken as a blanket assumption without considering the specific setting of the community under investigation.

### 5.2.2 Methodological innovations

Having established a fundamental commitment of the variationist approach to orderly heterogeneity, this section next discusses some key methodological principles that have been developed to best capture structured variation.

First, the variationist approach focuses on the collection of the ‘vernacular,’ i.e. a kind of spontaneous speech ‘reserved for intimate or casual situations,’ before ‘any efforts at (hyper-) correction or style shifting are made’ (Poplack, 1993, p.252). This means that researchers take speakers’ natural production in their everyday life as their prime source of investigation. As indicated in [Chapter 3](#), this is the type of data that CanVEC collected. I will not repeat the difficulties and advantages of this method here (see [Chapter 3](#), §3.2.1 instead), but the most important point to recall is that the ultimate aim of studying the vernacular is to ‘observe how people talk when they are not being observed’ (Labov, 1984, p.30).

Second, variationist methods pioneered the use of multivariate statistics for data analysis. This statistical method captures the orderly heterogeneity central to language use by modelling the simultaneous, multi-dimensional factors that impact on speakers’ choices and their regularities in the dataset (Tagliamonte & Baayen, 2012, p.12). This is based on the central belief of inherent variability: the application of any grammatical rule is **probabilistic** rather than categorical, and the presence or absence of certain features makes its application more or less likely. In other words, the method is not simply concerned with whether something occurs, but also with how often and where it occurs in order to identify what factors favour and disfavour its occurrence.

There are two main limitations of this method that we should be aware of. First, it is not possible to consider every factor that might influence a linguistic variable, i.e. we can only analyse the data based on a limited number of chosen factors. Second, statistical modelling depends on

the random distribution of the data and so statistical correlation does not necessarily amount to linguistic meaningfulness (Kiesling, 2011, p.24). There is no straightforward solution to these limitations; it is up to the analyst to ask the right question, to systematically interpret the patterns and to craft the most substantiated explanations for the data at hand (Tagliamonte, 2011, p.157).

Ultimately, successful execution of this statistical focus requires appropriate identification of the variable contexts and the possible conditioning factors to be coded for. This process is not random but must be grounded in linguistic knowledge of the variety (Wolfram, 1993, p.216). The next section therefore provides the necessary background to understand the variables that will be of central interest: the realisation of subjects, objects and copulas in Vietnamese.

## 5.3 Subjects, objects, and copulas in Vietnamese

### 5.3.1 Subject pronominal forms in Vietnamese

Vietnamese falls into the category of radical null subject languages (Biberauer, Holmberg, Roberts & Sheehan, 2010), i.e. a language that permits the omission of pronominal forms without verbal agreement of any kind (Thompson, 1965; Nguyen, 1997; Brunelle & Le, 2014). Examples (76)–(78) illustrate this system. Overt subject pronominal forms are highlighted in **boldface** and null subjects are represented by a  $\emptyset$  character.

- (76) a. **Con** đi đâu đấy?  
2SG go where there  
'Where are **you** going?'  
b. **Con** ra bưu-điện.  
1SG go-out post-office  
'I am going to the post-office.'
- (77) a.  $\emptyset$  Đi đâu đấy?  
go where there  
'Where are (**you**) going?'  
b.  $\emptyset$  Ra bưu-điện.  
go-out post-office  
'(**I**) am going to the post-office.'
- (78) a. **Cô-ấy** làm cho ai?  
3SG work for who  
'Who does **she** work for?'

- b. *Ø làm cho toà-đại-sứ Mỹ.*  
 work for embassy America  
 ‘(She) works for the American embassy.’

(Examples reproduced, with adapted glosses and translations, from Nguyen, 1997, pp.211–212)

As the data illustrates, Vietnamese subject pronominal forms can be expressed or unexpressed, across all grammatical persons.

The key fact that distinguishes Vietnamese from other radical pro-drop languages, however, is that anaphoric reference in spoken Vietnamese can be established not only by reduced pronominal forms but also by kinship terms and personal names (Ngo, 2006; Nguyen, 2018). Examples (79)–(81) demonstrate this system.<sup>76</sup> As subject pronominal forms are the focus, only pronominal forms in subject positions are in **boldface**.

- (79) a. Speaker A: **mày** qua Bến-Tre lần nào chưa? [Pronouns]  
 2SG cross Ben-Tre time which IMPERF  
 ‘Have **you** ever been to Ben-Tre?’

- b. Speaker B: **tao** đi hồi...  
 1SG go when  
 ‘I went when...’

- c. Speaker B: thì **mày** qua cầu là Ø tới Châu-Thành rồi.  
 then 2SG across bridge COMP arrive Chau-Thanh PERF  
 ‘(Once) **you**’ve crossed the bridge, (you)’ll arrive at Chau-Thanh.’

(Brunelle’s data, 2020, glosses and translations mine)

- (80) a. Speaker A: **chị** thích ăn món nào nhất? [Kin terms]  
 2SG.kin like eat dish which best  
 ‘Which dish do **you**-SISTER like the most?’

- b. Speaker B: **chị** thì thích ăn nhiều món của người à người Hoa á.  
 1SG.kin TOP like eat many dish POSS people DM people China DM  
 ‘I-SISTER like to eat many Chinese dishes.’

(Brunelle’s data, 2020, glosses and translations mine)

- (81) a. Speaker A: **Hiển** có giúp được chị việc này không? [Personal names]  
 2SG.name AFF help ASP.Acquired 1SG.kin business DEM NEG  
 ‘Could **you**-Hiên help me-SISTER with this?’

- b. Hiên: vâng, **em** giúp chị được  
 DM 1SG.kin help 2SG.kin ASP.Acquired  
 ‘Yes, I-SISTER can help you-SISTER.’

(Examples reproduced, with adapted glosses and translations, from Ngo, 2020, p.44)

<sup>76</sup>All examples deriving from Brunelle are from a series of natural recordings of southern Vietnamese conversations made by Marc Brunelle, as part of a grant from the Social Sciences and Humanities Research Council of Canada 435-2012-0468. The data was kindly made available to me upon request.

As we can see, in example (79), speakers use the pronouns *mày* and *tao* as 2SG and 1SG respectively. In (80), however, the kin term *chị* ‘sister’ is instead deployed as 1SG and 2SG (the grammatical role changes depending on who the speaker is, similarly to *con* in example (76) above). In the final example, the personal name *Hiển* in (81a) is used as a pronominal reference.

Of all the options, kin terms are the most commonly used to achieve anaphoric reference to individuals in speech (Chapter 3, §3.3.3). This is because personal pronouns and proper names in Vietnamese have been said to imply ‘a lack of deference and high degree of arrogance towards the addressee and/or third-party pronominal referent of superior age’ (Ngo, 2006, p.4). Vietnamese kin terms, on the other hand, show a ‘very deep concern for respect and good feeling’ among the interlocutors (Clark, 1988, p.21). As such, younger speakers must use kin terms rather than proper names and personal pronouns when speaking to or about their seniors. This is somewhat similar to the honorific system in Japanese (e.g. Hinds, 1975, 1983) but marks a striking difference to languages like English or Chinese, where pronouns are neutral.<sup>77</sup>

As a result of their loaded pragmatics, Vietnamese kin-term pronominal forms do not retain the literal meaning of kinship but instead index honorific information such as gender and age. In (80) for instance, *chị* does not project the core semantics of ‘sister’ but rather indexes speaker B’s gender and older age in comparison to speaker A. This rich indexicality of honorifics places extra pragmatic constraints on the occurrence of pronominal subjects in discourse: it is considered inappropriate for younger, or lower-social-status speakers to drop 1SG and 2SG pronominal forms in conversations (Nguyen, 1997; Pham, 2002; Do, Tran & Mai, 2018). In fact, as Ton (2018) shows in a corpus study, 98.5% of the drop of Vietnamese terms of address and reference in her data was accounted for by people of the same generation or an older generation talking to the younger generation (N=208), and only 1.5% in the reverse direction. A similar distribution was found with another set of data of 64 Vietnamese utterances collected by Le (2011) from natural conversations. (Kin) terms of address (2SG) in particular must be overtly expressed to appropriately convey due respect (Michaud & Brunelle, 2014). This variety-specific pragmatic norm is important to note, as it forms part of the conditioning factors that need to be accounted for in data modelling and analysis.

It is important to note, however, that this pragmatic constraint can be alleviated in a number of ways. Particularly, in spoken Vietnamese, the politeness marker *dạ* (utterance-initial), *vâng* (utterance-initial, Northern varieties) or *ạ* (utterance-final) are often used to offset 1SG pro-drop by younger generations.<sup>78</sup> This practice is demonstrated in (82):

<sup>77</sup>Song (2019, pp.126-129), however, challenges this assumption of Mandarin pronouns in recent work. Accordingly, he argued that the assumption of ‘neutral Mandarin pronouns’ is often based on a crude set of textbook data. When one examines real-life data more carefully, however, the category of pronoun in Mandarin turns out to subsume many more items with different semantic effects. See the discussion in Song (2019) for further details.

<sup>78</sup>For an extensive discussion of politeness markers in Vietnamese, see Vu (1997, 1999); Nguyen & Le (2013).

- (82) a. Speaker A: *em có hay lên Facebook chơi không?*  
 2SG.kin AFF often up Facebook play NEG  
 ‘Do you-YOUNGER hang out on Facebook often?’
- b. Speaker B: *ạ có.*  
 DM-POLITE AFF  
 ‘(I) do.’

(Brunelle’s data, 2020, glosses and translations mine)

Here, the 2SG pronominal form *em* produced by speaker A indicates that speaker B is younger than speaker A. Although speaker B dropped the 1SG pronominal form in her response to A, the construction is considered perfectly appropriate because the discourse marker *ạ* offsets the load for politeness. In research practice, this means that constructions with *ạ* and other politeness markers should be treated separately.

Returning to the context of this work, it should be further noted that the pragmatic effect of politeness markers only works for 1SG and not for 2SG pro-drop. In other words, the 2SG form as a term of address is strongly resistant to being dropped by younger/lower-socially ranked speakers, even in the presence of politeness markers of all kinds (Nguyen, 1997, p.211).

### 5.3.2 Object pronominal forms in Vietnamese

Similarly to subjects, object pronominal forms in Vietnamese can be expressed or unexpressed in a wide range of contexts, across all grammatical persons (e.g. Brunelle & Le, 2014; Phan & Lander, 2015). Examples (83)–(86) demonstrate this system.<sup>79</sup> Since object pronominal forms are the focus, only pronominal forms in object positions are in **boldface>**. The number in the square bracket is the number of intervening utterances not relevant to the point being made.

- (83) a. Speaker A: *tao giờ này tao đã biết bơi đâu.*  
 1SG time DEM 1SG PST know swim NEG  
 ‘I still don’t know how to swim now.’  
 [11]
- b. Speaker B: *rồi mấy thằng bạn mày rủ **Ø** đi tắm sông rồi sao?* [2SG]  
 then PL M friend 2SG invite go shower river then how  
 ‘But (if) your friends invite (**you**) to go swimming in the river, then what?’
- c. Speaker A: *tụi nó không có rủ **Ø*** [1SG]  
 PL 3 NEG AFF invite  
 ‘They don’t invite (**me**).’

<sup>79</sup>All the examples here are taken from Marc Brunelle’s recordings of colloquial Vietnamese. Note that object 2PL was not found in his corpus, and hence does not feature as an example here.

- (84) a. Speaker A: *rồi sau đó Ø mới ra ở nhà-trọ chung với nó hả?*  
 then after DEM just out at rental-home together with 3SG Q  
 ‘Then (you) moved in with **him/her** just after that?’
- b. Speaker B: *không, hồi trước Ø vô phòng trọ là gặp Ø rồi.* [3SG]  
 No time before enter room rent COMP meet PERF  
 ‘No, (I) had met (**him/her**) before when (I) checked in the rental room.’
- (85) a. Speaker A: *Ø có con rồi có-khi chồng bắt mình ở nhà [...],*  
 have child then maybe husband force 1PL at home  
 ‘Once (we) have children, (our) husbands might force **us** to stay at home.’
- b. *Ø hông có cho Ø đi làm.* [1PL]  
 NEG AFF let go work  
 ‘Not letting (**us**) to go to work.’
- (86) a. Speaker A: *bởi giờ tụi nó được hưởng Ø đó.*  
 that’s-why now PL 3 ASP.Acquired enjoy DM  
 ‘That’s why **they** can enjoy (all the inheritance) now.’
- b. *bà trả Ø rồi kiếm một mớ về quê nội.* [3PL]  
 3SG.F pay then earn one CLS return hometown paternal  
 ‘She (must) pay (**them**) and then earn something else to return home.’  
 (Brunelle’s data, 2020, glosses and translations mine)

Before discussing object omission further, however, a note of clarification is in order here on the transitivity of Vietnamese verbs: while object omission is restricted to certain verb types in languages like English (Allerton, 1975, 1982; Goldberg, 2001), this phenomenon is considerably more radical in discourse pro-drop languages like Vietnamese (Nguyen, 1997; Pham, 2002). In fact, any transitive verb may occur with or without an object, as long as it can be recovered from discourse. Analyses of radical pro-drop languages therefore often consider sentences such as (86b) to have a null direct object, whose reference is identified by a topic operator (e.g. Huang, 1984 for Chinese, Nakamura, 1991 for Japanese, Kim, 1989; O’Grady, Yamashita & Cho, 2008 for Korean).

Returning to the present discussion, despite similarly radical behaviours as to the grammatical environments where they can be dropped, what differentiates objects from subjects in Vietnamese is the fact that there is no culturally imposed pragmatic constraint on object drop in terms of grammatical person. In other words, younger speakers can drop objects in the same way that older speakers do, regardless of whether the object refers to older speakers or not. Example (87) illustrates.<sup>80</sup>

<sup>80</sup>These examples are part of some short recordings made in Ha Noi, Vietnam, by the Vietnamese Lexicography Centre (Vietlex). They were kindly made available to me upon request.

- (87) a. Speaker A: *Ø để em trả lần này.*  
 let 1SG.kin pay time DEM  
 ‘Let me-YOUNGER pay this time.’
- b. Speaker B: *Ừ thôi lần sau chị mời Ø*  
 DM DM time after 1SG.kin treat  
 ‘Okay I-OLDER will treat (you-YOUNGER) next time.’
- c. Speaker A: *rồi có gì em gọi Ø.*  
 then have anything 1SG.kin call  
 ‘Okay if there’s anything I’ll call (you-OLDER).’

(Vietlex data, glosses and translations mine)

In this example, we see that speaker A dropped the 2SG pronominal objects referring to her older interlocutor in (87c), just like her older interlocutor, speaker B, dropped the 2SG pronominal objects referring to her younger interlocutor (87b). In other words, in comparison to subjects, the pragmatic factors of interlocutor’s age and status are of little relevance to the expression of object pronominal forms in Vietnamese.

### 5.3.3 Copulas in Vietnamese

For copulas, variability exists between null versus overt realisation, particularly in the spoken variety. Some studies have suggested two types of overt copulas in Vietnamese: the copula-like conjunction *thì* and the ‘regular copula’ *là* (e.g. see Clark, 1992, 1996; Nguyen, 1997). Since the status of the copula-like *thì* is ambiguous, I focus on *là* only in this study.<sup>81</sup>

The copula *là* in Vietnamese is responsible for joining the subject and the predicate (or the topic and the comment) (Nguyen, 1997, p.118). In standard written Vietnamese, copula *là* is believed to be obligatory when it selects a nominal predicate but is omitted when it selects an adjectival predicate (see Nguyen, 1997, pp.85–86, Nguyen, Nguyen, Romary & Vu, 2004, p.4).<sup>82</sup>

<sup>81</sup>Specifically, the conjunction *thì* primarily functions as a topicaliser in several different structures, including [NP [*thì* Sentence]], [Subordinate Clause [*thì* Sentence]], [Sentence [*thì* Sentence]], and [NP/Sentence [*thì* Stative Verb]], where the segment preceding *thì* is the topicalised element (Clark, 1992). It is only in the [NP/Sentence [*thì* Stative Verb]] construction that *thì* behaves particularly like a copula with the stative verb describing the state of the event or the objects denoted by the NP/Sentence preceding *thì*. Although some authors have argued that *thì* and *là* differ little in meaning (e.g. Nguyen, 1957, 1975; Huffman & Tran, 2004), it is worth noting that *thì* does not always behave like a prototypical copula verb. In particular, it cannot replace *là* when the following predicate is an NP, neither can it be modified by an adverb, negativised, nor questioned like *là*. For a fuller description and examples of how *thì* functions, see Clark (1992, 1996).

<sup>82</sup>It is also worth noting that, in Vietnamese, *là* can also select a full clause as a predicate, as in (88) below. However, in these cases, *là* functions as a complementiser rather than a copula. These instances therefore lie beyond the scope of the discussion here.

- (88) *thế Ø bảo là Ø không biết là bà-nội ơ đi đâu*  
 so say COMP NEG know COMP grandma uh go where  
 ‘So (I) said that (I) don’t know where you are’

(Spoken Vietnamese reproduced, with adapted glosses and translations, from Ha, 2012, p.41)

In colloquial Vietnamese, however, this has been shown to vary (Clark, 1996; Wetzer, 2013). Examples (89)–(90) and (91)–(92) illustrate the variability with nominal and adjectival predicates respectively.

- (89) [...] *nhỏ đó mới là chảnh ty.* [NP-là-NP]  
 girl DEM then COP real Snake  
 ‘That girl **is** the real Snake.’ (in the context of the Chinese zodiac)
- (90) *mẹ mà y Ø thợ-may?* [NP-Ø-NP]  
 mother 2SG tailor  
 ‘(Is) your mum a tailor?’
- (91) *con-gái Hà-Nội rất là thanh-lịch, khéo-léo.* [NP-là-AdjP]  
 girl Ha-Noi INTSF COP elegant tactful  
 ‘Ha-Noi girls **are** very elegant and tactful.’
- (92) *nó Ø nhỏ-xíu.* [NP-Ø-AdjP]  
 3SG petite  
 ‘It (is) petite.’

(Brunelle’s data, 2020, glosses and translations mine)

As we can see, the copula *là* may or may not be expressed, regardless of whether the following predicate is an NP or an AdjP. This sets Vietnamese apart from other languages in the radical null subject languages group such as Mandarin, Japanese and Korean, which do not allow copula drop with nominal predicates. Instead, Vietnamese is more similar to Thai and other geographically adjacent languages in this regard (e.g. Wetzer, 2013, p.228).<sup>83</sup>

It is crucial to note, however, that while the realisation versus non-realisation of *là* in Vietnamese does not seem to have any consequences where the copula predicate is an NP, it is consequential for AdjP predicates. More specifically, a null copula Ø is the preferred option for AdjP, with the expressed variant reserved only for emphasis and assertiveness (Clark, 1996). In these cases, however, *là* must be accompanied by an intensifier (93), a perfective (94), or both. Without these discourse supporters, the occurrence of *là* in these examples is not acceptable (Diep, 2004, p.103).

- (93) *Gói hàng này là rất nặng* [COP-INTSF-Adj]  
 CLS goods DEM COP INTSF heavy  
 ‘This package (of goods) is very heavy.’
- (94) *Gói hàng này là nặng rồi* [COP-Adj-PERF]  
 CLS goods DEM COP heavy PERF  
 ‘This package (of goods) is already heavy.’

(Examples reproduced from Diep, 2004, p.103, glosses and translations mine)

<sup>83</sup> A visual contrast of the typology of copulas can be seen on the World Atlas of Language Structures, available at <https://wals.info/feature/120A#2/18.0/153.5>.

It should be noted, however, that this does not mean that copulas are obligatorily overt in order to achieve emphasis in Vietnamese. In fact, speakers can still emphasise or assert the attribute of something by placing the stress on the adjectival predicate without the realisation of the copula at all. Constructions such as (93) and (94) above, for example, are perfectly permissible even if the copula *là* is omitted. Whenever a copula is overtly realised in an AdjP, however, it must be accompanied by appropriate particles and/or intensifiers. Ultimately, in relation to overt/null realisation of copulas, the key point is that copula expression varies for both AdjP and NP predicate environments in Vietnamese.<sup>84</sup>

## 5.4 Previous studies on the realisation of subjects, objects, and copulas in a cross-generational context

In this section, I present a cross-linguistic overview of the way in which the realisation of subjects, objects, and copulas can vary in a cross-generational context. It should be noted that previous work on the distinction between null and overt copulas in a cross-generational context is particularly difficult to locate, and as such the majority of the space in the next section will be devoted to subjects and objects.

### 5.4.1 The transmission of subjects, objects, and copulas across generations

#### 5.4.1.1 Subjects

Despite having enjoyed the most research attention, findings regarding the cross-generational transmission of pronominal subject expression remain inconclusive. For example, in a recent variationist study investigating Spanish subject pronoun expression in New Mexico, a long-established contact community in the U.S, Torres Cacoullós & Travis (2018) compare the conditioning factors of modern Spanish in New Mexico with an earlier stage of the same variety. Their

<sup>84</sup>The only place where the realisation of *là* is obligatory is when the subject/topic is a clause and there is ambiguity as to whether the stative is applied to the whole clause, or to part of it, as demonstrated in (95).

- (95) a. *Anh làm là đúng.*  
 brother do BE BE.correct  
 'It is right for you to do it.'
- b. *Anh làm Ø đúng.*  
 brother do correct  
 'You do it correctly.'

(Reproduced from Nguyen, 1975, p.158)

In this example, the omission of *là* would change the POS of *đúng* from an Adjective modifying the preceding clause *anh làm* 'You do it' in (95a) into an adverb modifying the verb *làm* 'do' in (95b), thereby changing the meaning of the utterance.

results show evidence for continuity rather than change. In particular, all speakers demonstrate a robust distinction between null and overt forms, with the division being roughly 70:30 for null and overt subjects across both generations. Both generations also share the same factors conditioning null subjects, in the same direction of effects and relative strength of effects. In contrast, Otheguy, Zentella & Livert (2007) observe the opposite effect in their work on Spanish spoken in New York City (NYC). Analysing a corpus of 142 speakers from the six largest Spanish-speaking communities in NYC, Otheguy et al. (2007) show that although both groups maintain a high proportion of null subjects, Spanish speakers who arrive in NYC after the age of 16 and have been living there for less than six years produce a significantly higher rate of null subject pronouns than those who were born and raised in NYC (70% and 62% respectively). Otheguy et al. attribute this to the widespread bilingualism of the second generation (NYC-born), which is ‘concomitant with diminished levels of skills in, and frequency of use of, Spanish’ (p.795). This is to show that successful transmission of subject pronouns is variety-specific, as has been observed in a wide range of contradicting results for different contact varieties (cf. Backus, 2005 for Turkish, Bolonyai, 2000 for Hungarian, Lozano, 2006; Margaza & Bel, 2006 for Greek, Sorace & Filiaci, 2006 for Italian.)

In one of the most prominent comparative studies investigating null subjects cross-linguistically using identical variationist data and methods, Nagy (2015) also confirms this variety-specific tendency. Specifically, she conducted sociolinguistic interviews with 39 native speakers of Cantonese, Italian, and Russian spanning three generations in Toronto, Canada. Her results show that for Cantonese and Italian, there is no significant difference in the use of null subjects between those who were born in the homeland and those who were born in Toronto. Although some differences in raw rates emerge for heritage Cantonese, these differences disappear when the linguistic context is considered: first-generation Cantonese speakers happen to use more contexts that favour null subjects, i.e. cases in which the subject has already been introduced elsewhere. This significantly inflates their rates of null subjects.

For Russian, however, Nagy (2015) reports two cross-generational effects:

- (i) the hierarchy of grammatical person has been re-ordered across generations: **third person** > **second** > **first** for Gen 1 vs. **second** > **first** > **third person** for Gen 2, where factors closer to the left have a greater likelihood of being dropped; and
- (ii) while negation emerges as a significant predictor for null subjects in the second generation, it is not an existing predictor for first-generation speech.

Contrary to Otheguy et al. (2007), these differences correlate neither to the frequency of use of Russian nor to the ordering of the conditioning factors in English. Nagy (2015) takes this as evidence for internal cross-generational change in Russian, which is independent from contact.

In aggregate, the key fact that previous studies have highlighted is that cross-generational change in relation to null subjects does not follow a single pathway of change: change can be present in some varieties, but not in others.

#### 5.4.1.2 Objects

For null objects, similarly contradictory results have been reported, despite limited work being done on this variable. The most recent **and** comprehensive study is that of Rinke, Flores & Barbosa (2017), which examines Portuguese object omission in spontaneous production by second-generation European Portuguese-German bilingual speakers. Rinke et al. compare data from bilingual second-generation migrants against first-generation migrants and another two age-matched groups of monolingual speakers to separate internal change from contact effects. Their results show that the rate of null object use seems to pattern by age rather than by bilingualism. In particular, younger speakers in both bilingual and monolingual groups produce more null objects in their speech than the two older generations. Rinke et al. takes this as evidence for cross-generational change having taken place, albeit independently from contact.

Looking at differential object marking in heritage Spanish, Montrul & Sánchez-Walker (2013) presented a different result. Differential object marking in Spanish refers to a phenomenon where the speaker employs the overt preposition *a* to mark direct human accusative objects (e.g. *Perdí a mi hijo* ‘I lost my child’) and  $\emptyset$  to mark direct non-human direct objects (e.g. *Perdí  $\emptyset$  mis llaves* ‘I lost my keys’). Although differential object marking is not an obvious case of null versus overt object pronoun expression per se, Schwenter (2014) proposes a parallel between this phenomenon and the object marking system in Portuguese; specifically, Spanish direct objects typically occurring with accusative *a* encode the same features as overt pronominal direct objects in Portuguese, i.e. humanness, specificity and/or definiteness. In contrast, Spanish direct objects that occur without the *a* marking correspond to null direct objects in Portuguese. The realisation of Spanish differential object marking *a* can thus be seen as a proxy for overt pronominal objects for encoding similar features, while the omission of *a* signals a drop of a pronominal form. In the context of heritage Spanish in the United States, Montrul & Sánchez-Walker (2013) report that child and adult heritage speakers significantly drop differential object marking with animate and specific direct objects (~50–94%), just like first-generation immigrants. This tendency is in stark contrast with homeland speakers in Mexico, who display a very low rate of omission (~10%). In this sense, the rate of overt use here patterns by bilingualism rather than by age, a result that dif-

fers from Rinke et al. (2017). Similar to subjects then, evidence for or against cross-generational change of pronominal objects under contact remains elusive.

### 5.4.1.3 Copulas

For copulas, previous studies on cross-generational transmission are even more limited. In fact, while copulas have typically been cited as one of the most salient examples of language-contact influence, (given the prevalence of the null copula in the African diaspora; e.g. Bailey, 1966; Holm, 1984; Poplack & Tagliamonte, 1991; Rickford, 1997, 1998; Wolfram & Myrick, 2017), very little has been said about the distinction between null and overt forms in a cross-generational context.

The most explicit result to date can be found in Dannenberg's (2002) variationist study on null copulas in the Lumbee English vernacular in Robeson County, America. This dataset is extracted from 39 tape-recorded sociolinguistic interviews (19 Lumbee; 10 Anglo American; 10 African American) across three age groups: old (60+); middle (30–59); and young (10–29).<sup>85</sup> Results show no age effects for the Lumbee and African American groups, but report a difference for the Anglo American group. Specifically, the middle-aged Anglo American speakers strongly favour null copulas, while the young speakers strongly disfavour them. Dannenberg (2002) considers this evidence of a cross-generational change, indicating a possible shift among younger Anglo speakers towards a more standard English variety. As for motivations of change, Dannenberg (2002) attributes this to either the changing economic structure in the county or to the identity differentiation among the groups. Comparative work corroborating or challenging this result, however, is still rather limited. Further research is therefore needed to enable strong conclusions about the general behaviour of null and overt copulas in a cross-generational context.

### 5.4.1.4 Summary

The recurring theme from all of the studies reviewed thus far is that cross-generational change for subjects, objects, and copulas cannot be universally predicted. This is perhaps not surprising, given that the social conditions vary from one community to another, and the linguistic factors also differ between pairs of languages (e.g. Weinreich et al., 1968; Labov, 1972; Thomason & Kaufman, 1998; Muysken, 2000; McConvell, 2010; Trudgill, 2011, i.a.). The overview thus again highlights the need to situate the investigation in the specific sociolinguistic context of the vari-

---

<sup>85</sup>Robeson County, North Carolina, is located southeast of the state bordering South Carolina, America. The county is a tri-ethnic community that consists of approximately equal proportions of three ethnic groups: Native Americans (40% of the county population), African Americans (another 25%), and Anglo Americans (the remaining 35%). Since the 1700s, Robeson County has been inhabited simultaneously by these three ethnic groups (Dannenberg, 2002, p.357).

ety, as well as the community in which it is spoken. In the next section, I thus consider some Vietnamese-specific extra-linguistic factors for pronominal subjects and objects.

### 5.4.2 Pragmatic norms and cultural distance in language contact

In his seminal work, ‘Dynamics of Language Contact,’ Clyne (2003, p.215) states that ‘it is not always possible in this field to differentiate ‘language’ from ‘culture’ as a source of communicative behaviour.’ This is particularly true when it comes to the use of Vietnamese pronominal forms. As we have seen in §5.3.1, the pragmatic loads carried by different pronominal forms (including pronouns, kin terms, personal names, as well as their null variants) make them not only a linguistic but also a pragmatic and cultural instrument. Given that pronominal subjects and objects are the topics of interest in this chapter, relevant extra-linguistic considerations merit some further discussion here.

The first work looking at the nuances of Vietnamese pronominal forms in a contact setting was that of Tuc (2003), a study I previously discussed in [Chapter 4](#) in relation to the MLE. In this study, Tuc (2003) observes that the Vietnamese pronominal system is rather complex, and it can be unclear in contact situations which forms should be used. While kin terms are said to convey solidarity and respect, some pronouns such as *tao* ‘I’ and *mày* ‘you’ can be used to either express hostility or reinforce solidarity, depending on the relationship between the participants. These relationships are indirectly defined by existing social networks and social structures, which are not always as clearly defined in a diaspora setting as they are in the homeland. Thus, in order to avoid communication breakdown, speakers often prefer using English pronouns (Tuc, 2003) or dropping pronominal forms altogether (Nguyen, 2012). Tuc (2003), for example, cites the following case:

- (96) a. Speaker: *Cuốn sách này có cái cô này bị tạt a-xít [...]*  
 CLS book this has CLS Miss this PASS splash acid  
 ‘This book is about a girl who was attacked with acid [...].’
- b. Interviewer: *Sao, Ø bị tạt đánh ghen à?*  
 why PASS throw hit jealous PRT  
 ‘Why, is she hit because of a love affair?’
- c. Speaker: [...] **She’s** about seven or eight one time her dad *về nhà bắt-gặp mẹ she*  
 go home catch-up mother  
 with another man *nên ba she lấy a-xít tạt mẹ she nhưng-mà trúng she*  
 so father take acid throw mother but catch  
 ‘No, when she was about seven or eight years old, her dad came home one day to find her mother with another man. Her father attacked her mother with acid but it unfortunately ended up on the girl’s face.’  
 (Examples reproduced, with adapted glosses and translations, from Tuc, 2003, p.127)

In this example, Tuc (2003) explains that the speaker was telling a fictional story about a girl who was attacked with acid. In response, the interviewer asked if the girl was a victim of a love affair. This speculation about a love affair arises on account of the speaker having used the pronominal kin term *cô*, which indexes a young woman, at the beginning of the conversation. The speaker then realised the confusion and immediately corrected it to the English pronoun ‘she’ which covers all female referents regardless of age. As (96c) subsequently makes clear, the referent was only a seven- or eight-year-old girl. According to Tuc (2003), the switch to an English pronoun here was to ‘simplify’ the communication, and to avoid the risk of violating the Vietnamese regularities of correct pronominal forms. Similarly, Nguyen (2012) cites the example I give in (97). Here the speaker avoids using a 1SG self-reference pronominal form to index social relationships because the interlocutor (a taxi driver) is a stranger. For new encounters, it is often difficult for the speaker to assess the interlocutor’s age.<sup>86</sup>

- (97) Speaker: *Anh cho Ø tôi nhà-hát lớn*  
 2SG.kin.M let go Opera-House big  
 ‘Could you take (me) to the big Opera House?’

(Examples reproduced, with adapted glosses and translations, from Nguyen, 2012, p.134)

In the context of this work, since all CanVEC speakers chose their own interlocutors whom they knew well, cases of pro-drop due to speakers’ uncertainty like (97) are not relevant. What these examples together show, however, is how Vietnamese speakers deploy different pronominal forms in a contact scenario to serve different needs, as well as the prominent role that pragmatic considerations play in their decisions.

The salience of pronominal form-related pragmatic norms in contact scenarios has also been shown in a bilingual context, where Vietnamese kin terms are inserted as pronouns in an otherwise English context to achieve certain pragmatic functions. Specifically, Nguyen (2018) investigates a corpus of seven natural conversations of a parent-child dyad, and identifies consistent and frequent use of Vietnamese kin terms in place of English pronouns for self- and interlocutor-reference. The key finding is that most speakers cite the community norm as a reason for their choice of Vietnamese pronominal kin terms over English pronouns. In fact, despite some nuanced differences in speakers’ interpretation, both the first and the second generation treat Vietnamese pronominal kin terms as an indicator of their ‘Vietnamese-ness,’ and in retaining these linguistic forms, aim to retain the identity that is embedded in these items. First-generation speakers see this as a deliberate effort to ensure that the second generation pays ‘due respect’ to people in the community, while second-generation speakers see it simply as a way to ‘connect

<sup>86</sup>Note that the 2SG *anh* in this example does not imply its typical indexicality of an older male, but rather serves as a politeness honorific. Old speakers still address taxi drivers as *anh taxi*, or *bác tài* (lit. ‘uncle driver’) to pay respect to their professions.

better with other Vietnamese' (p.462). As such, this not only highlights the role of the pragmatic load embedded in Vietnamese pronominal forms, but also the speakers' awareness of the importance of these loads. In the context of this work, given that the omission of Vietnamese pronominal forms in monolingual Vietnamese is also known to index pragmatic implications (§5.3.1), this is a crucial point to note.

Having described the basic principles of the variationist framework and considered what previous studies have said about subjects, objects, and copulas across generations, I next apply these principles to analysing the CanVEC dataset.

## 5.5 Analysing CanVEC: Data coding and method

All the linguistic variables chosen in this study, i.e. pronominal subjects, pronominal objects and copulas, have two forms: an overt form and a null ( $\emptyset$ ) form. In order to probe the linguistic conditioning of null forms, the first step is to cross-tabulate their rates of occurrence against non-occurrence. This step is often referred to as identifying the 'envelope of variation,' i.e. the totality of situations where speakers have a choice between the variants (Labov, 1972, p.72). Although some researchers have taken a more restricted threshold of excluding only the absolute invariant environment (i.e. environments where subjects, objects, and copulas are categorically null or expressed), I instead favour the 'low-variability discount' approach, which excludes environments where variation is less than 5%, or greater than 95% (see Tagliamonte, 2006). According to Otheguy et al. (2007), the advantage of this less restrictive threshold is to avoid 'sterile discussions' around whether the very low variability of the environments is actually a case of zero variability, or whether they are in fact speakers' 'errors' (p.76). It hence follows that, when the variability is so minimal, 'it is best for the analyst, from a practical point of view, to proceed as if there were no variability at all' (ibid.). This predefined procedure is consistently applied to each of the dependent variables in the corpus.

### 5.5.1 Coding the dependent variables

I first extracted all Vietnamese monolingual finite clauses with speaker pseudonyms and time stamps onto an Excel spreadsheet. Each clause was subsequently marked for the presence or absence of an overt form for each variable. The marking process was a combination of automated retrieval (for expressed forms) and manual extraction (for null forms). In the case of automated retrieval, I searched each clause for the presence of a Vietnamese pronominal form or the copula

*là*.<sup>87</sup> Results were also manually verified line by line to ensure that no ‘right’ form in a ‘wrong’ position (e.g. *con* ‘I/you’ as a vocative instead of a subject or an object, etc., or *là* as part of a lexical compound instead of a copula) was extracted.<sup>88</sup>

In the case of completely manual retrieval, each remaining clause was judged as to whether there was a null form of subjects, objects, or copulas. For subjects and objects specifically, it should be recalled that Vietnamese pronominal forms are not part of a closed class system of function words, but are instead derived from a complex system of kin terms, personal pronouns or speakers’ names (§5.3.1). Since these three categories are all productively used as self-, interlocutor- or third-party references in the corpus, they all count towards ‘subject expression’ in Vietnamese.

In what follows, I will discuss specific cases that are further excluded for each variable.

### 5.5.1.1 Subjects

#### 5.5.1.1.1 Exclusion

First on the list of exclusions are Vietnamese set phrases that have been lexicalised in discourse. These include *thôi kệ* ‘(you/we) just ignore/leave (him/her/it)’ and *nếu mà nói là* ‘if (I/you) say that’ in the corpus (N=5, 0.1%). Common phrases such as *cảm ơn* ‘(I) thank you,’ *xin lỗi* ‘(I) am sorry’ are also excluded because although they can vary, they rarely do so in the data (see also Otheguy et al., 2007; Otheguy & Zentella, 2012; Torres Cacoullous & Travis, 2018).

Next are cases where the subject is ambiguous (N=24, 0.4%) as they are not codable in terms of linguistic factors such as person-number or coreferentiality. (98) represents an example.

- (98) a. Billy<sub>1</sub>: *quân-sư*    *á*    *ạ?*  
 agony-aunt    DM    DM  
 ‘Agony aunt?’
- b. Tyler<sub>2</sub>: *quân-sư*    *mà*    *quân-sư*    *ình-cảm*    *luôn-luôn*    *bị*    *ế.*  
 agony-aunt    but    agony-aunt    love    always    PASS    single  
 ‘The agony aunt who gives love advice is always single.’
- c.            That is a fact.
- d. Ellie<sub>1</sub>: *em*    *chưa*    *làm*    *quân-sư*    *cho*    *ai*    *bao-giờ*    *cả.*  
 1SG.kin    IMPERF    do    agony-aunt    for    anybody    ever    DM  
 ‘I haven’t acted as agony aunt to anybody.’

<sup>87</sup>A full list of searched forms can be seen in the forthcoming Table 5.5, §5.5.2.1.1.

<sup>88</sup>In order to fully automate this process, we need a parser that can determine the grammatical role of each instance. However, this is extremely complex to achieve on natural speech, especially on a low-resource language like Vietnamese. A trial run of several parsers in Vietnamese points to sub-par performance, making fully automatic extraction of overt forms a non-viable option.

- e.             $\emptyset$  *nói nhiều quá.*  
                  talk much INTSF  
                  ‘( $\emptyset$ ) talk too much.’

(Billy.Tyler.Ellie.0807, 18:18.8–18:31.9)

Here, the subject of *nói nhiều quá* ‘talk too much’ in (98e) is ambiguous. It is possible that Ellie is saying directly to both Billy and Tyler that they are talking too much (2PL), that one of them is talking too much (2SG), or that she is saying to one of them that the other person is talking too much (3SG). Yet another interpretation is that giving love advice to people involves (Ellie) talking too much, in which case the subject would be 1SG. As there are not enough clues from either the discourse or the syntax to determine which scenario is more likely, instances such as  $\emptyset$  in (98e) are excluded.

#### 5.5.1.1.2 Partial exclusion

Clauses with unintelligible tokens (marked as <V>) are dealt with next.<sup>89</sup> These clauses are included only if the unintelligible tokens do not affect the structural analysis of overt versus null pronominal subjects, as demonstrated in example (99).

- (99) a. Thomas<sub>1</sub>: *lúc đó anh nghèo.*  
                  period DEM 1SG.kin poor  
                  ‘I was poor at the time.’
- b.            *lúc đó là  $\emptyset$  chưa xài <V>.*  
                  period DEM COP IMPERF use  
                  ‘(I) didn’t use <V> at that time.’

(Max.Thomas.0823, 14:47.7-14:49.3)

In this instance, <V> is an unintelligible noun within a VP, and so has no effect on whether the clause has an overt or a null subject. It is clear from the remaining transcription that we have a null subject (1SG, recovered from discourse) in (99). Similar cases are included in the analysis.

In contrast, example (100) is ambiguous as to whether an overt or a null subject has been selected:

- (100) Theresa<sub>1</sub>: <V> *có cần me làm chả-giò không?*  
                  AFF need 1SG.kin make spring-roll NEG  
                  ‘Do(es) (X) need me to make spring rolls?’

(Theresa.Twee.0715, 25:48.9–25:50.2)

<sup>89</sup>Recall from the transcription convention in Chapter 3, §3.3.1.1, that when a token is unintelligible, it is marked as <X>, but if the transcriber has an idea of what language the token was produced in, the token would be marked <E> for English and <V> for Vietnamese, according to the transcriber’s ‘best guess.’ For our purpose, we do not consider clauses with <X> because it is not clear whether the token was English or Vietnamese, which in turn determines whether the clause was indeed monolingual. Similarly, we do not consider clauses marked with <E> because they either belong to the monolingual English subset, or the code-switching subset of the corpus.

In this example, the unrecognised word might be an overt subject for the verb *cần* ‘need,’ or an unintelligible word before a null subject. Accordingly, it is not possible to determine whether a null or an overt subject has occurred, and so these cases are excluded from the analysis (N=11, 0.2%).

Next, whenever repetition or repair (i.e. when speakers correct themselves) occurs (N=45, 0.8%), I treat the last form as the intended one. For example, in (101), I only count *con*<sub>2</sub> towards the totality of null vs. overt subject pronominal forms.

- (101) Quinn<sub>2</sub>: *tại-vì trên đó con<sub>1</sub> **con**<sub>2</sub> đói bụng quá,*  
 because up there 1SG.kin<sub>1</sub> 1SG.kin<sub>2</sub> hungry tummy INTSF  
 ‘Because I was really hungry over there.’

(Helen.Vivian.Quinn, 17:25.2–17:30.2)

In Vietnamese, 3SG *nó* can be used as either a neutral pronoun or as a (non-obligatory) expletive. Given that the study is only interested in referential subject pronouns, instances of 3SG expletive *nó* (102b)<sup>90</sup> are excluded (N=15, 0.3%).

- (102) a. Tim<sub>1</sub>: *sao ở nhà Ø không tập?*  
 why LOC home NEG exercise  
 ‘Why don’t (you) exercise at home?’

- b. Jess<sub>1</sub>: [...] *vì **nó** lạnh.*  
 because EXPL cold  
 ‘Because it’s cold.’

(Tim.Jess.Chloe.0705, 04:59.5–05:04.9)

When *nó* is used as a gender-neutral animate 3SG pronoun as in (103), however, it is counted (N=599).

- (103) Dany<sub>1</sub>: *thì nó viện cớ này cớ kia.*  
 then 3SG make-up excuse DEM excuse DEM  
 ‘He then made up different excuses.’

(Lami.Dany.0825, 01:39.8–01:41.7)

Similarly, Vietnamese *minh* can be used to refer to the non-specific 1PL ‘we’ (similarly to English generic 2SG ‘you’) (N=15) or to the specific 1PL ‘we,’ which includes the speaker and the interlocutor (N=249). The specificity of the referent makes a decisive difference as to whether or not they are admitted into the envelope of variation. Note that in this work, I make no distinction between ‘referents’ and ‘participants’; that is, the term ‘referent’ is used independently of grammatical person and can cover self- (1SG), interlocutor- (2SG) and third-party- (3SG) reference. In the context of the present discussion, consider example (104). The relevant pronominal subject *minh* is marked in **boldface**.

<sup>90</sup>See also Greco, Phan & Haegeman (2018) for a helpful discussion of *nó*.

- (104) a. Dany<sub>1</sub>: *họ phải có cái khuôn,*  
3PL must have CLS framework  
'They must have a framework'
- b. *cho mình biết,*  
let 2PL know  
'Letting us know'
- c. *là **mình** làm cái gì.*  
COMP 1PL do CLS what  
'What **we** are doing.'

(Brian.Dany.0812, 14:18.4–14:20.8)

In this example, Dany was commenting on the teacher's instructions for Brian's homework. As Dany does not attend schools anymore, *mình* in this instance does not refer specifically to Brian and Dany, but rather to all other students. In CanVEC, when the referent is non-specific, pronominal forms are categorically expressed and therefore lie outside the variable contexts (N=15, 0.3%).

By contrast, consider (105) in the following:

- (105) a. Ellie<sub>1</sub>: *em có cảm-giác là,*  
1SG.kin have feeling COMP  
'I have a feeling that.'
- b. *tối nay **mình** sẽ thức rất khuya.*  
evening DEM 1PL FUT stay-up INTSF late  
'**We** will stay up very late tonight.'

(Billy.Tyler.Ellie.0807, 04:47.5–04:51.9)

Here, it is clear from context that the 1PL *mình* refers specifically to Ellie and her interlocutors as people who would stay up late that night. When the referent is specific, there is enough variability in the corpus and so this kind of *mình* is included in the analysis.

Finally, since the present study is only concerned with subject pronominal forms, subject NPs such as *con-gái* 'girls' in (106) are not considered (N=722, 12.5%).

- (106) Dany<sub>1</sub>: *con-gái rất thích con-trai học giỏi hơn mình.*  
girl INTSF like boy study well more 1.REFL  
'Girls really like boys (who) study better than them.'

(Brian.Dany.0812, 21:01.3–21:05.6)

Taken together, [Table 5.1](#) summarises the special contexts that are either fully excluded or partially excluded from the envelope of variation for subjects.

SUBJECTS			
Linguistic context	Counted	N	%
Set phrases	x	5	0.1%
Subject NPs	x	722	12.5%
Ambiguous subjects	x	24	0.4%
Unintelligible tokens	*	11	0.2%
Repetition & Repair	*	45	0.8%
Expletive 3SG	*	15	0.3%
Generic pronouns	*	15	0.3%
<b>TOTAL</b>		837	14.6%

Table 5.1: Exclusions from the variable contexts for subjects in Vietnamese. Crosses (x) indicate full exclusion and asterisks (\*) signal partial exclusion as specified in the text.

### 5.5.1.2 Objects

#### 5.5.1.2.1 Exclusion

For objects, first to be excluded are clauses with intransitive verbs, i.e. those that cannot take a direct object (107). The Vietnamese Dictionary Vdict<sup>91</sup> is used as a source of reference. Specifically, I obtained a list of intransitive verbs from Vdict and automatically extracted clauses that contain them for exclusion.<sup>92</sup>

(107) a. Theresa<sub>1</sub>: *tối nay là Sabby sẽ ngủ trễ nè phải không?*  
 evening DEM COMP Sabby FUT sleep late DM right NEG  
 ‘Sabby will **sleep** late tonight, right?’

b. *Sabby tới hai giờ rưỡi Sabby mới thức dậy.*  
 Sabby until two hour half Sabby then wake-up DM  
 ‘Sabby only **woke up** at 2.30 (pm).’

(Theresa.Twee.0715, 12:45.3–12:53.0)

As previously noted in §5.3.2, the classification of ‘optionally transitive’ verb is not well-defined, as objects are freely dropped for most lexical verbs (Nguyen, 1997; Pham, 2002). There-

<sup>91</sup><https://vdict.com/>

<sup>92</sup>To explain why dictionaries were portrayed as unreliable in Chapter 3, but are used here to check transitivity, it is important to note that the purposes are different. In Chapter 3, we needed a means to distinguish a borrowing from a code-switch, the criteria for which depended on frequency and diffusion. As dictionaries are known to lag behind current usage of new words, they are unlikely to accurately reflect the status of a given lexical item. This problem is amplified in established bilingual communities away from the homeland, as the frequency or diffusion of any foreign word is expected to diverge from the monolingual community where data for traditional dictionaries are collected. For the purposes of identifying an envelope of variation, however, we need not establish the current usage status of a foreign word, but rather only ascertain its grammatical transitivity. This information is unlikely to change so quickly that traditional dictionaries cannot keep up. Furthermore, given the absence of better means such as descriptive work on the varieties under investigation, the use of reference dictionaries provides additional support for the researcher’s own grammatical judgements of individual verbs.

fore, I follow the protocol that verbs that are clearly intransitive such as *ngủ* ‘sleep’ or *thức* ‘wake up,’ as we see in (107), are straightforwardly excluded (N=201, 13%), while all other Vietnamese lexical verbs are considered more or less transitive and admitted into the envelope of variation. A complete list of Vietnamese intransitive verbs in the corpus can be found in [Appendix J](#).

Furthermore, as the current study is only concerned with pronominal direct objects, NP objects are not considered (N=60, 3.9%). Verbs that optionally subcategorise for a locative such as *về* ‘return’ in (108) are also excluded.

- (108) Quinn<sub>2</sub>: *rồi sau đó mình về nhà.*  
 then after DEM 1PL return home  
 ‘Then after that we go home.’

(Helen.Vivian.Quinn.0818, 05:37.1–05:48.2)

Specific constructions in the corpus with *muốn* ‘want’ and *cần* ‘need’ are similarly set aside (N=9, 0.6%). This is because these verbs take on a variety of complements, and so when a complement is dropped (109), it is not always straightforward to determine whether the dropped element is a pronominal object, an NP, or a VP.

- (109) a. Dany<sub>1</sub>: *chỉ có-thể vô làm được bác-sĩ đó,*  
 3SG can enter do ASP.Acquired doctor DM  
 ‘She could come and work as a doctor.’  
 b. *mà chỉ không có muốn Ø.*  
 but 3SG NEG AFF want  
 ‘But she doesn’t want Ø.’

(Brian.Dany.0812, 01:43.6–1:46.4)

In this case, the sentence could be that *chỉ không muốn điều đó* ‘she doesn’t want **it/that thing**,’ *chỉ không muốn làm bác sĩ* ‘she doesn’t want (to) **become a doctor**,’ or *chỉ không muốn rằng chỉ phải làm bác sĩ* ‘she doesn’t want **that she’d be a doctor**.’ Since we are only concerned with pronominal direct objects, these cases are discounted altogether.

In CanVEC, Vietnamese pronominal objects are categorically null when the object referent is inanimate (N=111, 7.3%), a phenomenon that has also been observed for Chinese (Yuan, 1997). Pronominal animate 3SG objects (110), on the other hand, are variable and therefore admitted into the envelope of variation.

- (110) a. Ellie<sub>1</sub>: *đây nè bây-giờ anh xem nó đi,*  
 here DM now 2SG.kin look-at 3SG IMP  
 ‘Here you look at **him**,’

- b. *bây-giờ dân mạng vẫn chê Ø.*  
 now citizen Internet still humiliate  
 ‘Even now netizens are still humiliating **(him)**.’

(Billy.Tyler.Ellie.0807, 21:07.2–21:23.7)

Finally, instances where the object has been topicalised in the same clause in Vietnamese are not considered (N=122, 8.1%). In these cases, the object is categorically ‘null’ in its normal post-verbal position in the corpus, as example (111) illustrates.

- (111) Tom<sub>1</sub>: *cô-á*đ**<sub>i</sub> *thì có mỗi-mình mà*đ**<sub>i</sub> *biết*  $\emptyset$ <sub>i</sub>.  
 3SG<sub>i</sub> TOP there only 2SG know  $\emptyset$ <sub>i</sub>  
 ‘As for her, only you know.’

(Tom.Henry.0809, 15:51.6–15:53.2)

### 5.5.1.2.2 Partial exclusion

Similar to the procedure for subjects, unintelligible clauses or those with unreliable cues are partially excluded, i.e. cases where the <V> token directly impedes the judgement of whether a pronominal object has been realised (N=2, 0.1%).

Some set phrases in Vietnamese are also excluded if they have been lexicalised to an extent of invariance in the corpus (N=7, 0.5%). This includes *để xem* ‘let’s see,’ *để xem thế nào* ‘let’s see how,’ *thôi kệ* ‘just ignore/leave (him/her/it),’ *làm ơn* ‘excuse me,’ and *cảm ơn* ‘thank you.’ One phrase—*tội-nghiệp* ‘feel sorry for/pity (you/him/her/them/us)’—varies considerably in relation to whether it takes an object, and so is included in the count. Consider (112).

- (112) a. Ellie<sub>1</sub>: *em tội-nghiệp hấn*,  
 1SG.kin feel-sorry-for 3SG  
 ‘I feel sorry (for) **him**,’  
 b. *em thấy tội-nghiệp*  $\emptyset$ .  
 1SG.kin feel sorry  
 ‘I feel sorry (for **him**).’

(Billy.Tyler.Ellie.0807, 21:18.2–21:22.1)

Table 5.2 summarises the instances that are either partially excluded or fully excluded from the envelope of variation for objects.

### 5.5.1.3 Copulas

For copulas, the usual protocols regarding unintelligible tokens, repetitions, and repairs were applied as they were for subjects and objects.

Recall from §5.3.3 that the copula *là* can also act as a complementiser, but these cases are beyond the scope of the discussion here. Copulas are thus counted only in cases where they select non-clausal predicates, as (113) and (114) illustrate.

OBJECTS			
Linguistic context	Counted	N	%
Set phrases	x	7	0.5%
Object NPs	x	60	3.9%
Verbs with locative complements	x	9	0.6%
Ambiguous objects	x	7	0.5%
Inanimate objects	x	111	7.3%
Topicalised objects in the same clause	x	122	8.1%
Intransitive Verbs	x	201	13%
Unintelligible tokens	*	2	0.1%
Repetition & Repair	*	1	0.1%
Generic pronouns	*	3	0.2%
<b>TOTAL</b>		523	34.3%

Table 5.2: Exclusions from the variable contexts for objects in Vietnamese. Crosses (x) indicate full exclusion and asterisks (\*) signal partial exclusion as specified in the text.

- (113) Penny<sub>2</sub>: *bà là người Xin-ga-po mà.*  
 3SG.F COP people Singapore DM  
 ‘She is Singaporean.’

(Penny.Marie.Rory.0912, 03:48.5–03:50.8)

- (114) Tom<sub>1</sub>: *hai-mươi phần-trăm là được rồi.*  
 twenty percent COP fine PERF  
 ‘Twenty percent is already good.’

(Tom.Henry.0725, 41:51.3–41:53.0)

Table 5.3 summarises the instances that are either partially excluded or fully excluded from the envelope of variation for copulas. This also concludes the coding process of the dependent variables.

COPULAS			
Linguistic context	Counted	N	%
<i>Là</i> as a complementiser	x	108	10%
Unintelligible tokens	*	2	0.2%
Repetition & Repair	*	1	0.1%
<b>TOTAL</b>		311	10.3%

Table 5.3: Exclusions from the variable contexts for copulas in Vietnamese. Crosses (x) indicate full exclusion and asterisks (\*) signal partial exclusion as specified in the text.

#### 5.5.1.4 Corpus distribution: CanVEC subjects, objects, and copulas across generations

Table 5.4 presents an overview of the dependent variable distribution in the corpus. As we can see, speakers produce more overt forms than null forms most of the time across all three variables. I will return to this fact in §5.7.2.3; for the purpose of what is being discussed here, the crucial point is that when null forms are being used, the first-generation speakers consistently produce higher rates than the second-generation speakers.

Dependent Variable	Gen 1		Gen 2	
	N	%	N	%
<b>SUBJECTS</b>	4126	100%	818	100%
Null	1311	<b>31.8%</b>	258	<b>31.5%</b>
Overt	2815	68.2%	560	68.5%
<b>OBJECTS</b>	608	100%	384	100%
Null	145	<b>23.8%</b>	52	<b>13.5%</b>
Overt	463	76.2%	332	86.5%
<b>COPULAS</b>	671	100%	327	100%
Null	83	<b>12.4%</b>	31	<b>9.5%</b>
Overt	588	87.6%	296	90.5%

Table 5.4: Cross-generational distribution of null vs. overt subjects, objects, and copulas

A Chi-square test reveals that the only statistically significant difference in rates is that of null objects ( $\chi^2 = 15.7$ ,  $p < 0.01$ ). The cross-generational difference for subjects and copulas is non-significant. However, as we previously saw in Nagy's study (2015) on Toronto Cantonese, the initial impression given by statistical difference may not always align with what is shown by linguistic patterns. In other words, the observed cross-generational differences for null objects is possibly a result of extra-linguistic, rather than linguistic factors (Bailey & Tillery, 2004; Hernández, 2009; Travis & Lindstrom, 2016). In contrast, for subjects and copulas, it is possible that while the overall rates of null forms remain constant across generations, the predictors conditioning their realisation in different contexts may be undergoing change. In fact, we will later see in §5.6 that although the identical rates of null subjects across generations give the impression of no variation, a multivariate analysis finds that this is by no means the case.

## 5.5.2 Coding the independent variables

To ascertain the true meaning of raw rates, we turn to multivariate analyses that consider the simultaneous effects of various conditioning factors. The first step in doing so is to code potential predictors for null subjects, null objects, and null copulas in the corpus. The independent

variables in this study include both linguistic and extra-linguistic predictors. The linguistic factors will be presented first and are specific to each variable, while the extra-linguistic factors are presented afterwards and apply to all variables.

### 5.5.2.1 Subjects

For subjects, three independent linguistic variables are selected: Person-Number, Clause Type and Coreferentiality. The selection of these variables is supported by both previous cross-linguistic work and by Vietnamese-specific facts.

#### 5.5.2.1.1 Person-Number

Grammatical person and number have consistently been presented as one of the strongest factors conditioning subject expression. More specifically, first-person has been found to be the most commonly realised subject pronoun in Spanish (Ranson, 1991; Bayley & Pease-Alvarez, 1997; Flores-Ferrán, 2002; Posio, 2015), European Portuguese (Barbosa, Duarte & Kato, 2005), and Mandarin Chinese (Jia & Bayley, 2002), while second- and third-person are the most frequent overt pronouns in other varieties such as Russian (Nagy, Aghdasi, Denis & Motut, 2011), Brazilian Portuguese (Barbosa et al., 2005), or Santomean Portuguese (Bouchard, 2018). Competing explanations have been put forward for these differences, but there has been no consensus. As Torres Cacoullos & Travis (2018) point out, while the greater expression of first-person singular has been attributed to the ‘egocentric nature of verbal communication’ (Silva-Corvalán & Enrique-Arias, 2017, p.184, translated by Torres Cacoullos & Travis, 2018, p.106), the opposite is just as applicable in different contexts. For instance, in Javanese, speakers’ general desire not to put themselves forward means that 1SG most strongly conditions the drop of a subject. In the context of Vietnamese, we have also seen that such norms are not even static, but vary depending on the interlocutor and their associated status (§5.3.1). The presence of these different discourse conventions means that, for grammatical person-number at least, instead of looking for absolute universals, we need to consider variety-specific patterns that are currently in play (Schroter, 2019, p.29). In the context of this work, Table 5.5 captures the coding scheme for all person-number subjects in the corpus.

As Table 5.5 shows, Vietnamese lacks a one-to-one mapping between many pronominal forms and grammatical person-number. This is again largely due to the kinship system that Vietnamese adopts, which enables the same form to change its referential values based on the discourse. For example, the same pronominal form *con* in (115) may be 2SG if uttered by speaker Tanner (115a), but becomes 1SG if uttered by interlocutor Nina (115b).

Vietnamese subjects/objects	Meaning	Code
<i>tui, tôi</i>	first-person gender-neutral	1SG
<i>mình</i>	first-person specific	1PL
<i>mày</i>	second-person gender-neutral	2SG
<i>hắn, nó</i>	younger gender-neutral	3SG
<i>họ, bọn họ</i>	third-person	3PL
<name>	proper name	1SG/2SG/3SG
<i>anh</i>	older M	1SG/2SG/3SG
<i>chị</i>	older F	1SG/2SG/3SG
<i>em</i>	younger gender-neutral	1SG/2SG/3SG
<i>con</i>	child	1SG/2SG/3SG
<i>ba</i>	father	1SG/2SG/3SG
<i>mẹ</i>	mother	1SG/2SG/3SG
<i>chú</i>	middle-aged male, younger than your own father	1SG/2SG/3SG
<i>bác</i>	middle-aged male, older than your own father	1SG/2SG/3SG
<i>bà</i>	grandmother	1SG/2SG/3SG
<i>bà ngoại</i>	maternal grandmother	1SG/2SG/3SG
<i>bà nội</i>	paternal grandmother	1SG/2SG/3SG
<i>ông</i>	grandfather	1SG/2SG/3SG
<i>ông ngoại</i>	maternal grandfather	1SG/2SG/3SG
<i>ông nội</i>	paternal grandfather	1SG/2SG/3SG
<i>cô, dì</i>	middle-aged female	1SG/2SG/3SG

Table 5.5: The coding scheme for Vietnamese person-number subjects in CanVEC

(115) a. Tanner<sub>1</sub>: *con gọi họ chưa?*  
 2SG.kin call 3PL IMPERF  
 ‘Have you<sub>-CHILD</sub> called them?’

b. Nina<sub>2</sub>: *con chưa gọi Ø.*  
 1SG.kin IMPERF call  
 ‘I<sub>-CHILD</sub> haven’t called (them).’

(Tanner.Nina.0609, 01:43.2–01:46.7)

The marking of person-number is therefore done by hand and relies entirely on the researcher’s interpretation of the whole discourse. This task is rather straightforward, thanks to abundant discourse cues provided by the conversational nature of the CanVEC dataset.

### 5.5.2.1.2 Clause Type

The next variable that is annotated is Clause Type. Cross-linguistically, Clause Type has been recognised as having an effect in a wide range of varieties such as English (e.g. Harvie, 1998),

Russian (e.g. Nagy et al., 2011), Spanish (e.g. Licerias & Díaz, 1999; Travis, 2007; Orozco, 2015), and Chinese (Jia & Bayley, 2002; Li, Chen & Chen, 2012). Clause Types are classified as Imperative, Interrogative, Declarative Main Clause, and Subordinate Clause. Example (116) illustrates this system. Note that (116e) is provided only for context, but not considered as it belongs to the non-monolingual subset, a topic of investigation in Chapter 4.

- (116) a. Chloe<sub>1</sub>: *con đừng có bật cái đó lên.* [Imperative]  
 2SG.kin NEG AFF turn CLS DEM up  
 ‘You do not turn that one on.’
- b. Tim<sub>1</sub>: *giờ con thấy trong người ra-sao?* [Interrogative]  
 now 2SG.kin see inside body how  
 ‘How are you feeling now?’
- c. Jess<sub>2</sub>: *con không có đau đầu lắm.* [Declarative Main]  
 1SG.kin NEG AFF hurt head INSTF  
 ‘My head does not hurt too much.’
- d. Chloe<sub>1</sub>: *tại-vì nếu-mà con bệnh,* [Subordinate]  
 because if 2SG.kin sick  
 ‘Because if you are sick,’
- e. *con chỉ nên ăn cái fruit không.* [Not considered]  
 2SG.kin just should eat CLS only  
 ‘You should just eat the fruit only.’

(Tim.Jess.Chloe.0705, 15:27.0–16:04.1)

### 5.5.2.1.3 Coreferentiality

The final, most complicated linguistic predictor to be coded for subjects is Coreferentiality. This is linked back to a widely established notion called Accessibility, which refers to the extent to which the referent is recoverable from discourse. Since the pioneering work of Givón (1983) on topic continuity, studies have repeatedly shown that the more ‘accessible’ the referents, the ‘less coding materials’ (less phonetic bulk) they require. Unexpressed pronominal forms belong to the category of fewest coding materials, and thus correspond to a more ‘accessible’ reference. By contrast, expressed pronominal forms are believed to occur more in contexts of less accessible references, primarily to fulfil the function of contrast and emphasis. This kind of effect has been discussed at length in Chafe (1994) and Payne (1997), and was subsequently reported for a wide range of languages, ranging from discourse pro-drop languages such as Mandarin (Li & Thompson, 1979; Christensen, 2000; Jia & Bayley, 2002; Li et al., 2012), Cantonese (Nagy et al., 2011), and Japanese (Lee & Yonezawa, 2008) to many other unrelated varieties such as Spanish (Butt & Benjamin, 2004; Torres Cacoullos & Travis, 2018), Portuguese (Paredes & Vera, 1993), Persian

(Haeri, 1998), Arabic (Owens, Dodsworth & Kohn, 2013), Italian and Russian (Nagy et al., 2011), Bislama (Vanuatu creole) and Tamambo (indigenous language of Malo island) (Meyerhoff, 2009), Finnish (Frascarelli, 2018) and English (Travis & Lindstrom, 2016).

In this study, Accessibility is defined in terms of Coreferentiality based on whether the subject of the current clause has the same referent as the subject of the previous clause (Yes vs. No).<sup>93</sup> This is regardless of whether the preceding clause was uttered by the same or a different speaker. Given that Coreferentiality is discourse-dependent, and discourse is co-constructed, it is appropriate that pronominal subjects mentioned by an interlocutor are also taken into account.

Same reference (i.e. ‘Yes’ ✓✓) is exemplified in (117) where the referent is the same across clauses, and switch reference (i.e. ‘No’ ✕✕) in (118), where the target subject differs from that of its preceding clause.

(117) a. Mina<sub>1</sub>: *mẹ cũng không biết luôn.* [Coreferential – Yes ✓✓]  
 1SG.kin also NEG know DM  
 ‘I also don’t know.’

b. Pete<sub>2</sub>: *mẹ đếm cũng được mà.*  
 2SG.kin count also ASP.Acquired but  
 ‘But you can count.’

(Mina.Pete.0906, 03:47.0–03:49.8)

(118) a. Brian<sub>1</sub>: *con này nó học giỏi lắm,* [Coreferential – No ✕✕]  
 girl DEM 3SG study well INTSF  
 ‘This girl **she**’s a very good student.’

b. *chị đừng lo.*  
 2SG.kin NEG worry  
 ‘**You** don’t worry.’

(Brian.Dany.0812, 20:55.3–20:58.1)

Furthermore, Coreferentiality might partially occur where the anaphor and its antecedent are in a whole-part relationship (as in (119)), or a part-whole relationship, (as in (120)).

(119) a. Brian<sub>1</sub>: *Ø<sub>1</sub> đang bạn thôi,* [Whole-part Coreferentiality ✓✓]  
 PROG friends just  
 ‘(We) are just being friends.’

<sup>93</sup>Note that although this previous mention might occur as subject, object, or another syntactic role, my working definition of Coreferentiality looks to the subject of the preceding clause only. There are two reasons for this. First, it is because cross-linguistically, topics are prototypically subjects. This tendency is also reflected in the corpus: an examination of a random 10% of CanVEC data shows that topic frequently coincides with the subject of the clause (93%, N=698/750). Second, in a recent study on Vietnamese pronominal realisations, Ngo (2019) found that overt pronominal forms are strongly favoured by structural parallelism between the grammatical roles of the target subjects and the antecedents (subject-subject, object-object), while non-parallelism results in mostly NPs. Given that pronominal forms are the focus in this study, it is appropriately practical to limit subject coreferentiality to only the preceding subject.

- b.  $\emptyset_2$  *chớ có làm gì với nó đâu chị Ti.*  
 DM AFF do anything with 3SG NEG sister Ti.  
 ‘(I) am not doing anything with her, sister Ti.’  
 (Theresa.Twee.0715, 10:14.3–10:17.9)
- (120) a. Taz<sub>1</sub>: *anh nói với nó đi,* [Part-whole Coreferentiality ✓✓]  
 2SG.kin talk to 3SG IMP  
 ‘You talk to him,’
- b. *rồi mình xử-lý.*  
 then 1PL handle  
 ‘and then we handle (it).’  
 (Tee.Taz.0808, 24:15.0–24:26.7)

As we can see in example (119), the target singular subject ( $\emptyset_2$ —‘I’) behaves as coreferential with parts of the previous plural mention ( $\emptyset_1$ —‘we’). Similarly, in example (120), the target plural subject (*mình*—‘we’) partly coindexes the previous singular subject (*anh*—‘you’). These cases, however, are extremely rare in the corpus (N=5/4,944 for subjects), and a separate treatment of partial coreferentiality turns out to be too fine-grained. As such, cases of partial coreferentiality were simply marked as if they were fully coreferential (✓✓).

### 5.5.2.2 Objects

Given that Vietnamese allows both subjects and objects to be null in similar linguistic environments (see §5.3.2), I coded the same independent linguistic variables (Person-Number, Clause Type, Coreferentiality) for objects as I did for subjects. While the coding for Person-Number and Clause Type is straightforwardly the same as it was for subjects, Coreferentiality, however, manifests in a slightly different way. In what follows, I thus first explain how Coreferentiality is coded for objects (§5.5.2.2.1) before discussing why I exclude another predictor that potentially conditions null objects: Animacy (§5.5.2.2.2).

#### 5.5.2.2.1 Coreferentiality for objects

Unlike subjects, the scope of Coreferentiality for objects extends beyond the previous mention in the same grammatical role of the preceding clause. This is because of the differences in the extent to which subjects and objects are linked to the topic of the sentence. In one of the earliest accounts of null arguments in discourse pro-drop languages, Huang (1984) argues that null arguments (both subjects and objects) are identified by a null sentence topic, which is in turn grammatically linked to a discourse topic. What distinguishes objects from subjects, however, is the fact that contrary to null subjects which may occur in embedded clauses (121a), objects cannot do so in these positions (121b).

- (121) a. *Zhangsan<sub>i</sub>, ta<sub>i</sub> shuo e<sub>i</sub> mei kanjian Lisi.* [Chinese]  
 Zhangsan he say no see Lisi  
 ‘Zhangsan<sub>i</sub>, he<sub>i</sub> said that [he<sub>i</sub>] didn’t see Lisi.’
- b. \**Zhangsan<sub>i</sub>, ta<sub>i</sub> shuo Lisi mei kanjian e<sub>i</sub>.*  
 Zhangsan he say Lisi no see  
 ‘Zhangsan<sub>i</sub>, he<sub>i</sub> said that Lisi didn’t see [him<sub>i</sub>].’

(Reproduced from Huang, 1984, p.558)

According to Huang, example (121a) is grammatical in Chinese because an empty embedded subject can be identified by the matrix subject (i.e. ‘Zhangsan’ in this case). In contrast, example (121b) is ungrammatical because objects are not bound to matrix subjects. Instead, an empty object has to be identified by its closest nominal element, which is an empty topic. Huang (1984) takes this as evidence that objects in discourse pro-drop languages are inherently licensed by the topic.

Given that topicality is significant, and most topics coincide with subjects, it is not justified to limit object coreferentiality to objects only. Consequently, I extend the scope of Coreferentiality for objects to include pronominal subjects (122), as well as objects (123) and topicalised objects (124) in the immediately preceding clause. Coreferentiality also counts for previous mention by an interlocutor, as previously set out for subjects.

Similar to the illustration for subjects, same referent (i.e. ‘Yes’ for coreferentiality) is marked with ✓✓ and switch reference (i.e. ‘No’ for coreferentiality) is marked with ✕✕. The target object is in line (b.), and its antecedent is in line (a.), both are highlighted in **boldface**.

- (122) a. Dany<sub>1</sub>: *ch<sub>i</sub> thì **ch<sub>i</sub>** biết nó quá rồi,* [Antecedent]  
 1SG.kin then 1SG.kin know 3SG INTSF PERF  
 ‘As for me, I already know him very well.’

- b. *mà nó không biết **ch<sub>i</sub>**.* [Coreferenced with previous SUB ✓✓]  
 but 3SG NEG know 1SG.kin  
 ‘But he does not know **me**.’

(Lami.Dany.0825, 08:42.9–08:45.9)

- (123) a. Dany<sub>1</sub>: *ngày xưa **ch<sub>i</sub>** dạy nó,* [Antecedent]  
 day ancient 1SG.kin teach 3SG  
 ‘I taught **him** back in the old days.’

- b. *ch<sub>i</sub> chăm-sóc **Ø**,* [Coreferenced with previous OBJ ✓✓]  
 1SG.kin look-after  
 ‘I looked after (**him**).’

(Lami.Dany.0825, 21:21.7–21:23.1)

- (124) a. Billy<sub>1</sub>: *chị nó<sub>i</sub> mình gặp Ø<sub>i</sub> rồi mà.* [Antecedent]  
 girl DEM 1PL meet PERF DM  
 ‘That girl, we have met.’
- b. Tyler<sub>2</sub>: *anh chưa gặp Ø.* [Coreferenced with previous TOP OBJ ✓✓]  
 1SG.kin IMPERF meet  
 ‘I haven’t met (her).’  
 (Billy.Tyler.Ellie.0807, 01:19.1–01:21.3)
- (125) a. Theresa<sub>1</sub>: *ba-me dẫn nó tới hoài mấy lần,*  
 1PL.kin lead 3SG to frequently several times  
 ‘We took him there several times,’
- b. *mà hông khi-nào Ø gặp được bà hết.* [Non-coreferential ✕✕]  
 but NEG when meet ASP.Acquired her DM  
 ‘But (we) could never meet her.’  
 (Theresa.Twee.0715, 10:14.3–10:17.9)

Note that the topicalised object in the same clause is not included in the variable context, as the post-verbal slot, usually occupied by the object, is categorically empty (§5.5.1.2). However, they are considered for the purpose of coreferentiality with the object of the following clause as it determines the topic. In (124), for example, the null object in Billy’s utterance is not counted as a dependent variable, but the null object in Tyler’s is. Tyler’s null object is also further marked as coreferential with the topicalised object *chị nó* ‘she’ in Billy’s preceding clause, even though *chị nó* lies outside the variable context.

### 5.5.2.2.2 A note on Animacy

Another factor that has been specifically singled out as a strong predictor for object realisation is Animacy (Landa, 1995; Yuan, 1997; Choi, 2000; Colantoni, 2002; Schwenter, 2006; Meyerhoff, 2009; Schwenter, 2014; Lecanda & Schwenter, 2017). Despite this, however, Animacy is not coded in this study. This is because, similar to what has been reported for Chinese (Yuan, 1997), 3SG inanimate object pronouns are categorically absent in the corpus; therefore they are not admitted into the envelope of variation.<sup>94</sup> Example (126) illustrates a case in point.

- (126) a. Tim<sub>1</sub>: *với-lại cái heater nhà mình bị gì nữa,* [Inanimate antecedent]  
 and CLS home 1PL PSV something more/also  
 ‘And our heater also has some problems,’

<sup>94</sup>While Animacy might be relevant for subjects, it is confined to third-person only (first- and second-person are animate by default). Thus, in order to tease out the effects of Animacy proper, we must study third-person subjects separately. This is beyond the scope of this work and left for future studies.

- b. *ngày-mai* Ø *kêu người-ta sửa* Ø [Null anaphor]  
 tomorrow call people repair  
 ‘(I) will call people to repair (it) tomorrow.’

(Tim.Jess.Chloe.0705, 16:48.3–17:00.2)

Here, the antecedent of the pronominal object Ø in (126b) is the heater mentioned in (126a), which is inanimate. In these cases, the pronominal object is categorically null in CanVEC.

### 5.5.2.3 Copulas

For copulas, two of the strongest predictors that have repeatedly been found to condition null copulas are the Predicate Type (i.e. the grammatical category of the following constituent) and the Subject Type (a pronominal form vs. an NP). In this section, I will discuss these two variables respectively.

#### 5.5.2.3.1 Predicate Type

Cross-linguistically, a general tendency that has been found is that in languages that allow both nominal and adjectival copular complements, adjectivals are expected to be more prone to copula deletion than nominals (e.g. Labov, Cohen, Robins & Lewis, 1968; Labov, 1969; Baugh, 1980; Poplack & Tagliamonte, 1991; Walker & Meechan, 1999; Walker, 2000; Walker & Meyerhoff, 2006; Geeslin & Guijarro-Fuentes, 2008; Adamou, 2013; Sanchez-Alonso, 2018; Adamou, De Pascale, García-Márkina & Padure, 2019). In the context of CanVEC, we only consider copulas followed by an NP or an AdjP. This is because finer distinctions (such as locatives or CP) are not justified by the data distribution. In fact, together, locatives and CPs selected by a copula (both null and overt) only account for slightly more than 1% of the dataset (N=11/909).

Example (127) below demonstrates the coding output for Predicate Type. In this example, copulas and their corresponding translations are in **boldface**. The following predicates are underlined, while the number in the square bracket is the number of intervening clauses not relevant to the point being made.

- (127) a. Thomas<sub>1</sub>: *cho-nên kiểu sống khá là phóng-khoáng* [AdjP predicate]  
 so style living quite COP liberal  
 ‘So the lifestyle **is** quite liberal,’

- b. *nó Ø xô-bô lắm.* [AdjP predicate]  
 EXPL chaotic INTSF  
 ‘It (is) quite chaotic.’

[9]

- c. *rồi nó là cái lịch-sử của anh.* [NP predicate]  
 then EXPL COP CLS history POSS 1SG.kin  
 ‘So that is my history.’

(Max.Thomas.0823, 01:54.7–02:20.3)

### 5.5.2.3.2 Subject Type

The second predictor to be coded for copulas is Subject Type. This is a factor that has been identified mainly from the wealth of work on copula BE in AAVE (e.g. Rickford, 2006; Czinglar, Katicic, Kähler & Schaner-Wolles, 2008; Mobaraki, Vainikka & Young-Scholten, 2008; Sharma & Rickford, 2009; Kautzsch, 2012). The general observation is that pronouns prefer deletion (or contractions) of copulas while NP subjects prefer full forms.<sup>95</sup> Accordingly, I made a distinction between NP subjects (such as 127a) and pronominal subjects (127b and 127c) in marking subject type for copulas.

Having coded the independent linguistic variables, I next consider the extra-linguistic factors that potentially affect these variables of interest.

### 5.5.2.4 Extra-linguistic factors

Data for extra-linguistic factors was extracted from the questionnaire (Chapter 3, §3.2.2) documenting speakers’ linguistic and social information. All information collected from the questionnaire was first imported into an Excel spreadsheet. Note that unless there is a good reason to do so, not all recorded factors should be included in the model for statistical analysis (Kiesling, 2011; Tagliamonte, 2011). For example, it was found that ‘Age of acquisition’ and ‘Language taught at schools’ highly correlates with ‘Generation,’ and these two are therefore excluded. ‘Speaker’s occupation,’ ‘Level of education,’ ‘Caregivers’ primary language,’ ‘Attitudes to code-switching’ and ‘Self-reported behaviours of code-switching’ are also left aside. What remains are ‘Age,’ ‘Gender,’ ‘Primary language of the social network,’ self-assessed ‘Proficiency’ in each language, ‘Attitude’ towards each language, and ‘Speakers’ ethnic orientation’—all of which have been shown to play various roles in language variation (e.g. Labov, 1966, 1972; Trudgill, 1974; Labov, 1984; Milroy & Milroy, 1992; Chambers, 1995; Johnstone & Kiesling, 2008; Kiesling, 2009; Nguyen, 2015). Lan-

<sup>95</sup>Although Walker & Meyerhoff (2006) suggested that the subject type effect is more of a phonological effect, this account has been seriously challenged. For the Hamilton variety of English (Bequia island), Walker & Meyerhoff (2006) found that both pronouns and NPs ending in a vowel favoured a null copula, while NPs ending in a consonant disfavoured one. They take this as evidence for the phonological effects rather than the subject type itself. This conclusion, however, has already been met with criticism regarding some counter-evidence in their own dataset, as well as several problematic statistical assumptions (see Rickford, 2006 for a full critique). Furthermore, as Labov (1969) also shows in his study on AAVE, while there are fewer full forms after NPs that end with vowels than those that end with consonants, they still exceed full forms occurring after pronouns. In other words, the fact that pronouns end with vowels accounts for some, but not all of the effects on copula contraction and deletion. As such, while phonological environments may or may not be a factor, the effects of subject type are well-established.

guage attitude in particular is recorded in the form of a score in the range of 1-5 as previously reported (Chapter 2, §2.4.3 & §2.4.4).

Two further variables were coded specifically for subjects: ‘**Interlocutor’s Age**’ and ‘**Interlocutor’s Generation**.’ This is because, based on what we know about the honorific indexicality of Vietnamese pronominal forms (§5.3.1), pragmatic constraints are likely to have an effect. Although we may never be able to conclusively define situational ‘respect’ or ‘politeness,’ the obvious factors that play a role here are the age gap between speakers and their respective social statuses. This pragmatic constraint is thus operationalised as ‘Interlocutor’s Age’ and ‘Interlocutor’s Generation.’ Note that although clauses with politeness markers merit separate consideration (§5.3.1), they only account for a very small number in CanVEC (N=3, <0.01%) and are therefore not further analysed.

### 5.5.2.5 Summary

In this section, I have laid out the protocol upon which both linguistic and extra-linguistic independent variables are selected and coded for the realisation of Vietnamese subjects, objects, and copulas in CanVEC. Table 5.6 captures all the variables that were coded for as well as what was excluded and included for each variable.

## 5.5.3 Statistical modelling: Rbrul mixed-effects

Having completed the coding process, I submitted the output to Rbrul, a gold-standard statistical tool that was specifically designed to account for linguistic variation. In this section, I explain the core assumptions of Rbrul, as well as the final modelling process.

### 5.5.3.1 Rbrul explained

In order to understand the advantages of Rbrul, it is first important to note that before Rbrul, traditional regression modelling in variationist studies often treated individual data points as independent.<sup>96</sup> This is particularly problematic given that most sociolinguistic studies involve a finite number of speakers, each producing a different number of data points (this is also the case for the CanVEC data). The data thus already violates the assumption of independent observations. As Tagliamonte & Baayen (2012, p.143) point out, as soon as a given individual contributes

<sup>96</sup>Labov’s ‘fourth floor’ study on rhoticity in New York City (Chapter 2, §2.3.1) is again a good example of this. Each speaker in this study was asked a question that prompted them to utter the phrase ‘fourth floor’ so that the production of /r/ could be assessed. As each speaker produced exactly one example of the dependent variable, the assumed independence of observations matched the reality of the dataset. Most sociolinguistic studies, however, are not specifically designed this way.

Dependent variables				Independent extra-linguistic variables	
Property	Subjects	Objects	Copulas	Predictor	Included
Unintelligible tokens	*	*	x	Generation	✓
Repetition & Repair	*	*	x	Gender	✓
Expletive 3SG	*	*	–	Speaker's Age	✓
Generic pronouns	*	*	–	Language of social network	✓
Set phrases	x	x	–	Ethnic orientation	✓
Subject NPs	x	–	–	Vietnamese proficiency	✓
Ambiguous subjects	x	–	–	English proficiency	✓
Object NPs	–	x	–	Vietnamese lang. attitude	✓
Ambiguous objects	–	x	–	English lang. attitude	✓
Inanimate objects	–	x	–	Interlocutor's Age	*
Topicalised objects in the same clause	–	x	–	Interlocutor's Generation	*
Intransitive Verbs	–	x	–	Age of acquisition	x
<i>Là</i> as a complementiser	–	–	x	Language taught at schools	x
				Speakers' Occupation	x
				Level of education	x
				Caregivers' Primary language	x
				Attitudes to code-switching	x
				Code-switching self-reported behaviours	x
Independent linguistic variables				Legend	
Predictor	Subjects	Objects	Copulas	✓	inclusion
Person-Number	✓	✓	–	*	partial inclusion
Clause Type	✓	✓	–	x	non inclusion
Coreferentiality	✓	✓	–	–	non-applicable values
Animacy	–	x	–		
Predicate Type	–	–	✓		
Subject Type	–	–	✓		

Table 5.6: An overview of the coding scheme for subjects, objects and copulas

more than one observation, 'that individual becomes a source of variation that should be brought into the statistical model.'

The first advantage of Rbrul is its capacity to account for this individual 'source of variation.' It does this using the generalised linear *glmer* function (Bates, Mächler, Bolker & Walker, 2015), which both incorporates and differentiates between two types of predictor: fixed effects and random effects. Fixed effects are predictors that comprise a small numbers of variants that are intended to be replicable in other studies, while random effects are factors drawn from a larger population that are not usually replicable (Johnson, 2009, p.365). For example, a mixed-effects model would treat gender as a fixed effect, given the traditionally assumed two genders (Male/Female), but speakers as a random effect as it is highly unlikely that the same speakers will participate in all other studies. According to Johnson (2009, p.363), failure to factor in speaker

random effects can lead to gross overestimation of the significance of effects, returning statistically significant outputs that are likely a combination of chance and individual variation. The solution to this is to recognise the variable nature of input probability, allowing speakers to differ randomly without skewing the overall significance of the investigated effects.<sup>97</sup>

Second, Rbrul can model interaction terms between independent variables, optionally as part of the same automatic procedure identifying significant main effects. Historically, analyses only used a simple main-effect logistic regression model, which means predictors are considered to be completely independent from each other. For example, under simple main-effect modelling, a speaker's generation classification (Gen 1 or Gen 2) in CanVEC would be treated as separate from the linguistic conditioning of their expression of subjects (e.g. person-number of the token involved), without being able to consider how they might interact and create a joint effect on the dependent variable. As we will see in §5.6, however, this is a limiting assumption, as generation classification is in fact not totally independent from the person-number property of the (un)expressed subjects. Had we ignored this possible interaction, we might have a different conclusion in relation to cross-generational variation in the community.<sup>98</sup>

### 5.5.3.2 Rbrul modelling

To test for cross-generational differences, I include a fixed effect of Generation and interaction terms of Generation by each linguistic predictor. If any of these interaction terms emerges as significant, the inference is that cross-generational changes of some kind have taken place. All of the other predictors are also included to likewise establish which factors significantly condition the realisation of null elements in Vietnamese.

<sup>97</sup>Rbrul ensures this by not fitting a parameter around each individual's data, but only treating an effect as significant if the factor 'is strong enough to rise above the inter-speaker variation' (Johnson, 2009, p.365). Where extreme individual variation is found and chance creates 'the appearance of external effects,' Rbrul raises its standard accordingly. This means that no single outlier is responsible for the reported significant effect. While this approach is conservative and may mean that fewer significant results are detected, we can be much more certain that the effects returned truly are significant (Hay, 2011, p.212). Another point worth mentioning is that accounting for individual random effects also means that 'different internal constraints' of individual speakers are taken into consideration (Johnson, 2009, p.374), effectively addressing the criticism that individual agency is often disregarded in quantitative analyses using social categories. This is a remnant of the debate surrounding 'community grammar' in the 1970s, see Kay & McDaniel (1979); Sankoff & Labov (1979) for further details.

<sup>98</sup>On a broader scale, possible interactions are also inherent in many sociolinguistic data sets (e.g. see Mooney, 2018 for interactions between sex and syllable type, or Tagliamonte & Baayen, 2012 for a summary of work on interactions between age and sex). These interactions should be appropriately accounted for. Although previous studies mainly use simple main effects models (largely because the VARBRUL series could only handle interactions with difficulty, or require researchers to 'manually' split the data and run parallel regressions, see e.g. Paolillo, 2002 for more), it has been repeatedly shown that models with interaction terms almost always improve the overall goodness of fit (Johnson, 2009, p.381).

The initial run of the model included all factors in order to identify which emerged as the most important, and whether there was any problem of multicollinearity.<sup>99</sup> Further steps were also taken following Tagliamonte's (2012) recommendation to check for correlation by cross-tabulation and monitoring the regression.

As a result, the models were iteratively updated two times between the initial run and the final version. The first change includes the removal of 'Vietnamese language attitude,' 'English language attitude,' and self-reported 'Vietnamese proficiency' and 'English proficiency,' which are collinear with speakers' 'Ethnic orientation.'<sup>100</sup> The second change involves 'Interlocutor's Age,' which was initially coded as continuous, but later collapsed into a binary factor: Older/Younger (in relation to the speaker themselves). This is because while it is important for a **speaker's** age to be modelled as a continuous variable to avoid arbitrarily drawing a line in the population, it is not as essential for the **interlocutor's** age. In fact, the only element that has pragmatic relevance here is whether the interlocutor is deemed older or younger than the speaker. Accordingly, simplifying the variants for 'Interlocutor's Age' is desirable and can speed up the modelling process. Table 5.7 lists all the predictors and factors that are included in the model.

## 5.6 Results

Within the variationist approach, the working hypothesis is that 'competing variants will occur at greater or lesser rates depending on the features that constitute the context' (Poplack, 2001, p.405). Favouring factors are therefore identified via co-occurrence patterns within the envelope of variation. For example, we may predict that if null pronominal subjects are a grammatical device coding 'more accessible' referents (§5.5.1.1), they would occur at a higher than average rate in coreferential contexts. This means that in multivariate analysis, we expect that this linguistic sub-context will favour the occurrence of null subjects.

<sup>99</sup>Multicollinearity refers to situations where two or more independent variables correlate. For example, all but one Gen 2 speaker in CanVEC scored themselves the maximum mark 4 (Confident in extended conversations) for English (all the other Gen 2 rated themselves 3—Fairly confident). Similarly, almost every Gen 1 speaker rated themselves 4 for Vietnamese, with only two rating themselves 3. Speakers who score higher on language attitude towards Vietnamese also tend to identify themselves as 'more Vietnamese' in their 'Ethnic orientation,' and similarly for English. As generalised linear models make the assumption that independent variables are not collinear, this often creates 'unsolvable computational problems' (Tagliamonte & Baayen, 2012, p.163). Even if the program successfully returns an output, the log-odds or factor weights reported are highly unstable due to the inherent difficulties with factoring out collinearity effects.

<sup>100</sup>When the model fits the data in the step-wise regression, each single factor is added one at a time (step-up) or eliminated one at a time (step-down) to find the best fit. However, collinearities in the data result in a mismatch between the step-up and step-down. This is a problem mixed-effects modelling currently cannot deal with. I thus followed Rbrul's advice against including all the factors that interact with each other in the model (Johnson, 2009; Mooney, 2018), and removed self-assessed 'Vietnamese language attitude' and 'Vietnamese language proficiency'. This choice is made on the basis that 'Generation' and 'Ethnic orientation' arguably provide more precise data about the speakers than their self-assessment on relative concepts such as language proficiency and language attitude.

Linguistic factors			Extra-linguistic factors			
Dependent variable	Predictor	Factor	Predictor	Factor		
Subjects	Person x number	1SG	Speaker*	Speaker's unique ID		
		2SG	Generation	Gen 1		
		3SG		Gen 2		
		1PL	Gender	Male		
		2PL†		Female		
		3PL	Age**	Speaker's Age		
	Coreferentiality	Yes	Primary language of the social network	Vietnamese		
		No		English		
	Clause Type	Imperative	Main	Both		
				Subordinate	Ethnic orientation	More Australian
					More Vietnamese	
				Interogative	Neutral	
Objects	Person x number	1SG	Interlocutor's Age***	> the speaker		
		2SG		< the speaker		
		3SG	Interlocutor's Generation***	Gen 1		
		1PL		Gen 2		
		2PL†	<b>Interaction terms</b>			
		3PL				
	Coreferentiality	Yes	<b>Dependent variable</b>	<b>Factor</b>		
		No	Subjects	Generation x Person-Number		
	Clause Type	Imperative	Main	Generation x Coreferentiality		
				Subordinate	Generation x Clause Type	
					Interogative	Objects
				Generation x Coreferentiality		
Copulas	Predicate Type	Adjective Phrase	Generation x Clause Type			
			Noun Phrase	Copulas	Generation x Predicate Type	
Subject Type	Pronominal	Generation x Subject Type				
		NP				

† Only 5 tokens in CanVEC and therefore excluded from statistical analysis

\* Random intercept

\*\* Continuous variable

\*\*\* Included in the model for subjects only

Table 5.7: Factors included in Rbrul modelling

Table 5.8 presents the results of the three independent mixed-effects logistic regression models for our three variables of interest.<sup>101</sup> In this table, significant factors are listed alongside a factor weight (FW), while non-significant factors are listed at the bottom. The value of the factor weight ranges from 0 to 1 and indicates the **direction of effects**. Specifically, a value above 0.5 (in **boldface**) denotes a positive correlation, meaning the factor is more likely to facilitate null tokens, while a value below 0.5 denotes a negative correlation, meaning the factor is more likely to disfavour null tokens. The percentage in the second column indicates the proportion of null tokens for each factor while the third column reports the total number of clauses where each factor is applicable. Also included are the ranges of each predictor, which indicate the **size of effects**: the higher the value, the stronger the effect a given predictor has on the realisation of null tokens.

We can see from Table 5.8 that the prediction for Coreferentiality is not borne out for null subjects, but it is for null objects in the corpus. Specifically, null objects are favoured in contexts where the referent has already occurred in the preceding clause. ‘Language of the social network’ also comes out as a significant predictor for objects: speakers who use Vietnamese as the primary language are most likely to drop objects, followed by those using a mix of both, and least favoured by speakers whose primary language within the social network is English. For copulas, the only significant predictor is Predicate Type, with adjectival predicates favouring null copulas and nominal predicates disfavouring them. Some of these findings are particularly noteworthy, and will be returned to in detail in §5.7.2. For now, the key result for null objects and copulas is that neither Generation nor Generation crossing any linguistic predictors were selected as significant. This suggests a cross-generational stability rather than change.

For subjects, however, we see that other than Clause Type having an effect, there is a significant interaction between Generation and Person-Number. More specifically, the hierarchy of the constraints has been reordered across generations: while first-person pronominal subjects (both SG and PL) most strongly favour the realisation of null subjects in the first generation, they least favour null subjects in the second generation. Instead, second-generation speakers are most likely to drop 2SG subjects, as exemplified in (128).

- (128) a. Tressie<sub>2</sub>: Ø    coi    Ø ð    dàu?  
           (2SG) watch    LOC where  
           ‘Where did (**you**) watch (it)?’

<sup>101</sup>The variable rule analysis calculates the overall probability (i.e. the ‘input’ reported at the top of the table), which is a basic probability of applying the variable rule to the dataset. This number is calculated in Rbrul by averaging out all the predicted values of the proportion of null vs. overt tokens in each cell, i.e. the baseline upon which the model builds its prediction and selects predictors that significantly contribute to the model’s goodness of fit.

<b>SUBJECTS</b>		<b>N=1,569/4,944, Input: 0.29, Overall rate: 31.7%</b>		
<b>Generation x Person-Number</b>	<b>FW</b>	<b>%</b>	<b>N</b>	
Gen 1 x 1SG	<b>0.67</b>	36.8%	1166	
Gen 1 x 1PL	<b>0.54</b>	31.6%	258	
Gen 1 x 3PL	0.48	30.6%	121	
Gen 1 x 2SG	0.47	27.7%	1361	
Gen 1 x 3SG	0.45	11.5%	1220	
Gen 2 x 2SG	<b>0.59</b>	39.4%	257	
Gen 2 x 3SG	<b>0.55</b>	37.4%	183	
Gen 2 x 3PL	0.47	30.0%	30	
Gen 2 x 1SG	0.45	23.0%	313	
Gen 2 x 1PL	0.34	8.6%	35	
Range	33			
<b>Clause Type</b>				
Imperative	<b>0.62</b>	61.3%	193	
Interrogative	<b>0.59</b>	36.3%	966	
Declarative Main	0.42	30.2%	3097	
Subordinate	0.39	22.9%	688	
Range	23			
<i>Non-significant predictors: Person-number, Coreferentiality, Generation, Gender, Age, Primary language of the social network, Ethnic orientation, Interlocutor's Age, Interlocutor's Generation, Generation x Clause Type, Generation x Coreferentiality</i>				
<b>OBJECTS</b>		<b>N=197/992, Input: 0.17, Overall rate: 19.8%</b>		
<b>Coreferentiality</b>	<b>FW</b>	<b>%</b>	<b>N</b>	
Yes	<b>0.72</b>	36.6%	519	
No	0.29	10.0%	473	
Range	43			
<b>Social network lang.</b>				
Vietnamese	<b>0.71</b>	28.1%	371	
Both	<b>0.63</b>	20.9%	539	
English	0.30	2.9%	82	
Range	41			
<i>Non-significant predictors: Person-Number, Clause Type, Generation, Gender, Age, Ethnic orientation, Generation x Person-Number, Generation x Clause Type, Generation x Coreferentiality</i>				
<b>COPULAS</b>		<b>N=114/998, Input: 0.19, Overall rate: 11.5%</b>		
<b>Predicate Type</b>	<b>FW</b>	<b>%</b>	<b>N</b>	
Adjective Phrase	<b>0.83</b>	64.8%	552	
Noun Phrase	0.17	3.8%	446	
Range	66			
<i>Non-significant predictors: Subject Type, Generation, Gender, Age, Primary language of the social network, Ethnic orientation, Generation x Subject Type, Generation x Predicate Type</i>				

Table 5.8: Mixed effects model for Vietnamese null subjects, objects, and copulas

- b. Luna<sub>1</sub>: Ø    coi    Ø với Alana.  
 (1SG) watch    with Alana  
 ‘(I) watched (it) with Alana.’

(Luna.Tressie.0901, 06:26.6–06:30.5)

Taken together, the results show that cross-generational effects are only observable for null pronominal subjects in the Canberra Vietnamese community. In contrast, second-generation speakers show no significant divergence from their first-generation counterparts when it comes to null objects and null copulas. This contradicts what we previously saw in terms of percentages of null forms across generations (Table 5.4, §5.5.1.4), and again highlights the importance of looking beyond the initial impression given by raw rates.

In the next section, I consider the implications of these findings and what they mean in a broader context of language change and variation.

## 5.7 Discussion: Heritage language in the community

Within the variationist approach, it is typically the case that when a linguistic predictor such as ‘Person-Number’ is selected as significant (with reordered factor ranks), the conclusion is that a cross-generational change in speakers’ ‘grammar’ or competence has occurred. For Vietnamese, however, we saw in §5.3.1 that the expression of Person-Number of pronominal subjects encodes culturally loaded information, and so the concept of a ‘change in grammar’ is not so straightforward. The primary aim of the present discussion is thus first to consider the role of pragmatic norms in explaining the observed cross-generational variation for null subjects (§5.7.1). Having done so, I then move beyond variation to account for the stability observed in the results for null objects and null copulas, i.e. the linguistic and extra-linguistic factors that condition both first and second generations alike (§5.7.2).

### 5.7.1 Cross-generational variation: Traces of community bricolage

#### 5.7.1.1 The peculiar direction of effects for null subjects

The first finding that needs to be accounted for is the specific direction of effects for Person-Number in null subjects. This is because while the result for cross-generational variability itself is not surprising, the specific direction of effects contradicts the expectation. Specifically, although we expected the second generation’s terms of address (2SG) to be more likely to be overt than the first generation’s (§5.3.1), we observed the opposite trend in our results. The significance of this result needs to be considered in the context that the overt realisation of 2SG is virtually non-negotiable in terms of conveying respect in Vietnamese: unlike 1SG, the 2SG form as a term of

address still cannot be dropped by younger/lower-socially ranked speakers, even in the presence of politeness markers of all kinds (§5.3.1).

Situating this in the context of the Canberra Vietnamese community, I interpret this as a result of cultural assimilation into the Australian society. This is first because, despite maintaining a high level of cultural identity and cohesion, most speakers in Canberra work in the Australian public service, have other high-skilled jobs, or are pursuing formal education (Chapter 2, §2.3). This suggests a high level of integration into mainstream social life. Second, results from the background questionnaire show that most speakers (90%, N=40/45) identify themselves as ‘Both Vietnamese and Australian’ (Chapter 3, §3.2.2). This dual sense of identity in particular indicates at least some orientation towards the Australian cultural values, which, among others, include the ‘spirit of egalitarianism,’ embracing equality and social fairness (Kapferer & Morris, 2003; Thompson & Stannard, 2008).<sup>102</sup> In adapting to the majority’s community norms, one may speculate that speakers are consciously or subconsciously doing the opposite of what is expected in the heritage language (i.e. expressing 2SG towards younger speakers instead of dropping it) as a form of hyper-correction to offset the perceived lack of equality indexed in the Vietnamese information structure. This potential explanation can be tested in future work, when comparable data from homeland speakers and other Vietnamese diasporas becomes available.<sup>103</sup>

In the context of this work, nonetheless, the seemingly culturally integrated practice at issue can be construed as a form of community bricolage (in the sense of Eckert, 2004, §5.2.1.2). For the first-generation speakers, overt 2SG directed at younger speakers is possibly a way to identify themselves as ‘modern’ Vietnamese who treat those socially ‘lesser’ than themselves more equally, thereby distancing themselves from the homeland speakers.<sup>104</sup> This is particularly conceivable, considering the fact that many first-generation Vietnamese Australians still generally feel some form of distance or difficult emotions towards the ‘Communists’ in the homeland (Chapter 2, §2.2). As first-generation speakers also serve as input for second-generation speakers, it then

<sup>102</sup>The Australian values have also been formally documented in the Australian Values Statement by the Australian Government Department of Home Affairs. To see the full statement, visit <https://immi.homeaffairs.gov.au/help-support/meeting-our-requirements/australian-values>.

<sup>103</sup>Henriëtte Hendriks (p.c.) points out that given that speakers avoid kinship terms with unknown interlocutors, it could be that this system of kinship terms is becoming less readily usable for these speakers. In other words, they may use the null forms because they do not know what the appropriate form would be. However, remember that all CanVEC speakers chose their own interlocutors whom they knew well (Chapter 3, §3.2.1), and therefore cases of pro-drop due to speakers’ uncertainty does not seem a likely explanation here (§5.4.2).

<sup>104</sup>This is not only restricted to the cross-generational conversations in the corpus (N=16/23), but also applies to conversations between Gen 1 speakers only (N=6/23). This is because in all of the conversations in CanVEC, there is an age difference between speakers (Chapter 2, Table 2.2). This means that even if both speakers belong to the same generation, one speaker is always considered more socially ‘respectable’ than the other. Furthermore, it should be noted that although Generation does not necessarily correlate with age in CanVEC, in all of the conversations, the Gen 2 interlocutor is always younger than the Gen 1 speaker. In this sense, the majority of Gen 1 speakers are speaking to younger interlocutors in the corpus (N=20/28). For those Gen 1 who speak to older Gen 1 interlocutors, their overt realisation of 2SG is already expected (§5.3.1).

follows that the pragmatic norms of explicitly expressing subjects towards older interlocutors might never have been properly transmitted and therefore acquired by the second generation.

In the event that the pragmatic norm was actually transmitted despite the circumstances,<sup>105</sup> this account of bricolage may still apply. In particular, it is probable that the bricolage was innovated by second-generation speakers, and that by frequently dropping 2SG directed at older speakers, these younger speakers are trying to reject the Vietnamese social hierarchy entrenched in the language, thereby establishing a more equal relationship with the older generation.

In either case, the distancing behaviour observed here is consistent with the argument from some previous studies in Australia (cf. e.g. Clyne (2003) for heritage Dutch and German), but may seem to be at odds with Nguyen's (2018) study on Vietnamese kin terms in Australia. Specifically, Nguyen (2018) found that speakers consistently voice the desire to remain close to Vietnamese pragmatic norms for terms of reference and address. This difference in findings, however, likely stems from the fact that the sample in Nguyen (2018) is significantly smaller in size, sourced from a wider range of locations in Australia (as opposed to the Canberra Vietnamese community exclusively), and constrained only to the domain of parent-child conversations. Furthermore, Nguyen (2018) investigated pronominal kin term expression as a single word insertion in speakers' mixed speech, rather than in their monolingual Vietnamese. Data may thus not be comparable and speakers may have different norms for different language combinations. It is also possible that such norms are fluid, continually negotiated, reformed and developed. In any case, this complexity of competing norms signals the need for further research.

### 5.7.1.2 Inter-speaker variability

Although the variationist framework often prioritises community patterns over those of individuals, I made a case in §5.2.1.2 that the role of individual speakers must also be duly accounted for. With this in mind, Figure 5.1 and Figure 5.2 respectively capture inter-speaker variability in relation to null subjects among first- and second-generation speakers. Note that while only the percentages are reported here to make the general patterns clear, the raw counts per speaker per grammatical person-number can be found in Appendix K. As this appendix shows, all the speakers within the same generation produce comparable numbers of tokens of null subjects.

Setting aside the specific differences in percentages reported in Figure 5.1 for a moment, I would like to highlight the fact that the preference for dropped first-person subjects (represented by orange) holds across all 28 first-generation speakers in CanVEC. Specifically, the first-person

<sup>105</sup>It is not uncommon in the community that different standards apply to different generations (e.g. social protocols for subject drop, as we have seen in §5.3.1). Accordingly, first-generation speakers may decide that while they can do as they wish, it is important that younger speakers remain respectful at all times when they speak Vietnamese. In this case, the pragmatic norms of when to express a pronominal subject would be transmitted.

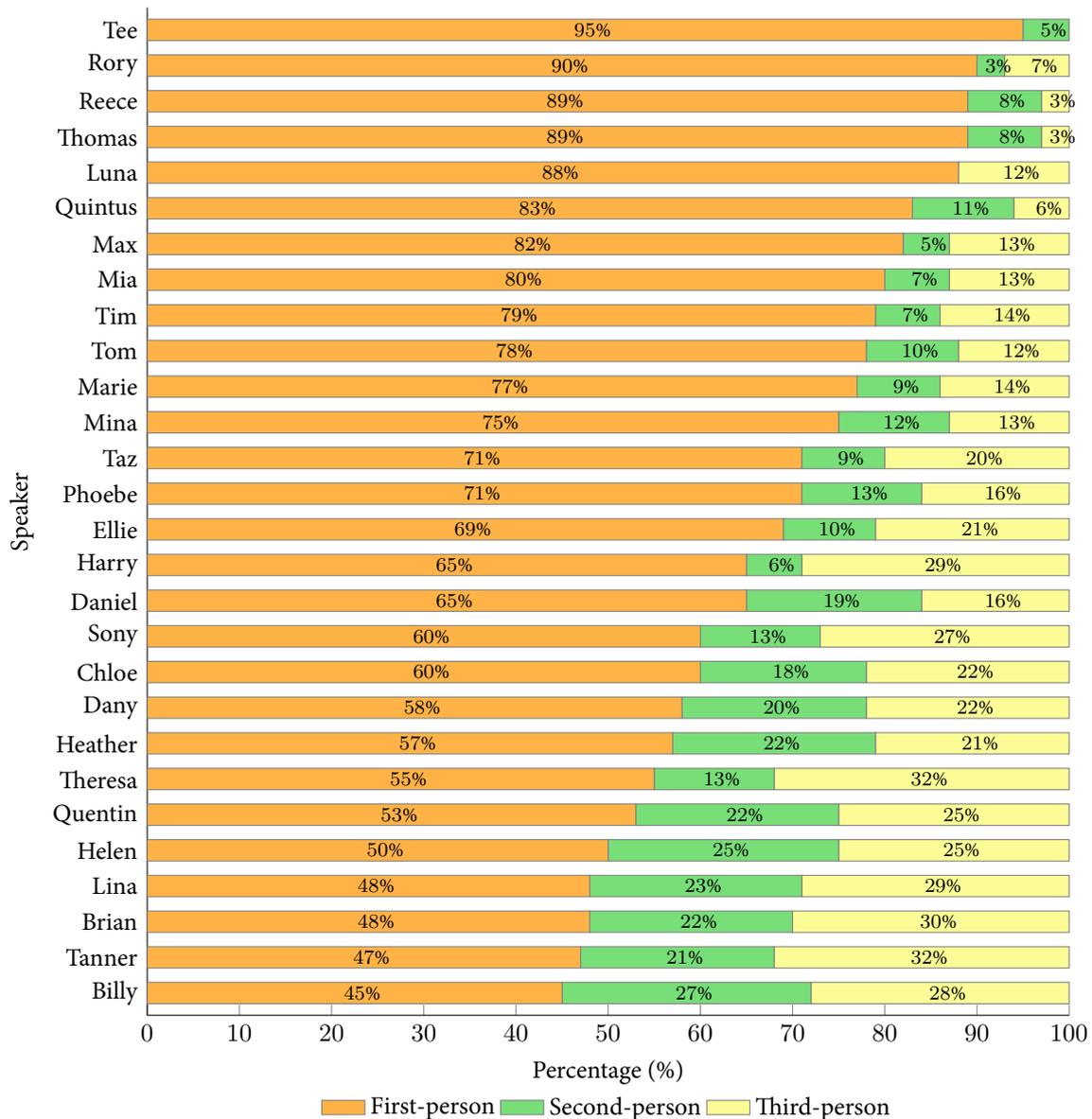


Figure 5.1: The distributional pattern of the first generation's Vietnamese null subjects by grammatical person

consistently accounts for the highest proportion of unexpressed subjects. It is crucial to note that this tendency is not a byproduct of skewed data distribution, as first-person pronominal forms (both null and overt) only account for just over one third of the total number of subject pronominal forms produced by first-generation speakers (35%,  $N=1,424/4,126$ ).

Similarly, the pattern is strikingly consistent for the second generation, albeit with two exceptions. Specifically, Figure 5.2 shows that second-person null subjects account for the highest number of unexpressed subjects for most speakers in the corpus. This result is pronounced, given the fact that in total, second-generation pronominal forms only make up 31% of the second generation's production (N=258/818).<sup>106</sup>

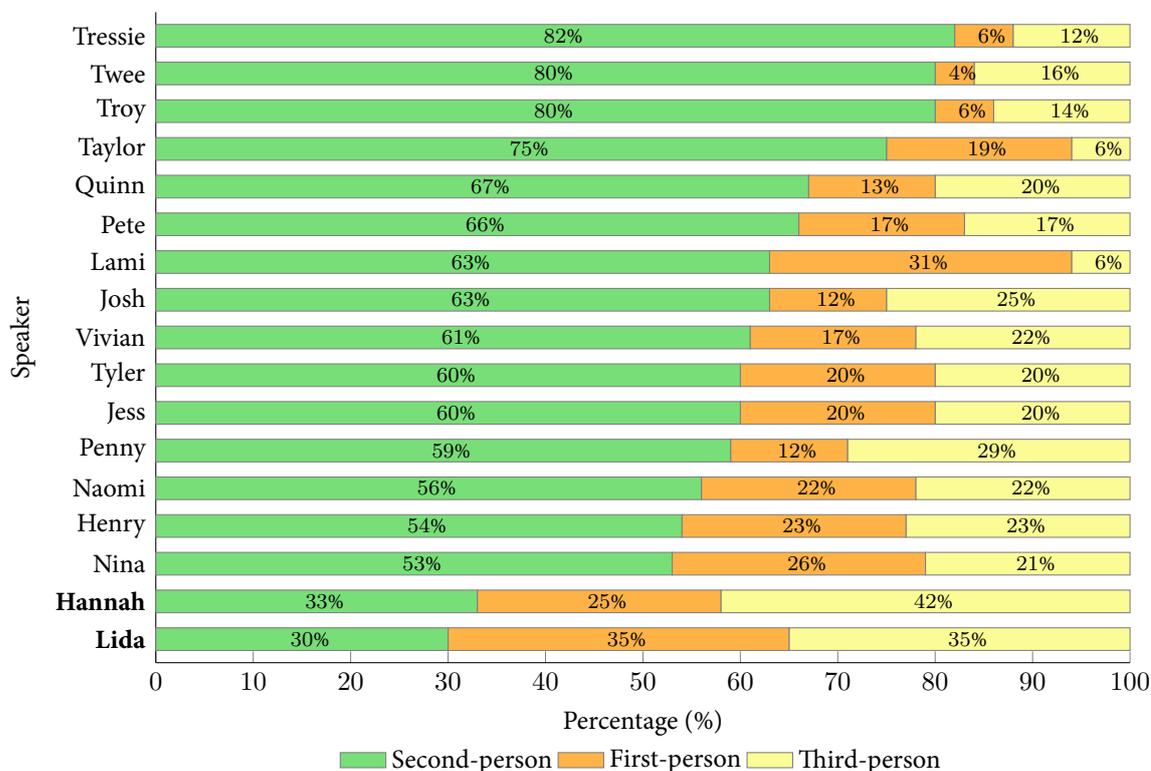


Figure 5.2: The distributional pattern of the second generation's Vietnamese null subjects by grammatical person

We also notice from Figure 5.2, however, that Hannah and Lida display a much lower proportion of second-person null subjects (33% and 30% respectively) than the rest (53% and above), with an almost even distribution across all null subjects. While their divergent behaviour may be explained by the topic of the conversation or other discourse factors, it is worth noting that both Hannah and Lida participate in the **only** CanVEC conversation where a first-generation speaker is **not** present. Given that Hannah and Lida do not noticeably diverge from any other second-generation speakers in terms of social and linguistic background (questionnaire, see Chapter 3, §3.2.2 and Appendix E), this fact about their conversation configuration is potentially significant. Although it is difficult to make a conclusive generalisation based on the only conversation between two second-generation speakers in the corpus, this observation possibly points to the

<sup>106</sup>It is the first-person forms that account for the largest proportion of the second generation's production (N=348/818, 43%). A detailed break-down of the total forms was previously shown in Table 5.8.

role of the interlocutor's generation membership.<sup>107</sup> In this sense, the finding here provides further potential support for the analysis of purposeful bricolage among second-generation speakers: since the drop of 2SG subjects is deployed to establish a more equal dynamic with the older generation, this pattern is not observable in conversations where the older generation is not involved. Further investigation of conversations between second-generation speakers only would thus be a useful focus for future research.

Returning to the specific percentages observed in each generation, we see that in fact, there is a fair amount of variance between speakers in terms of rates. For the first generation, the range is between 45-95% for their most preferred person subject drop (i.e. first-person), while for the second generation, this number is between 30-82% (i.e. for second-person). Such variance is typical of natural data in sociolinguistics, and has thus again justified the decision to use Rbrul mixed-effects modelling—a method that effectively factors in individual variation—so that no outliers are single-handedly responsible for the output. It is likely that these large ranges of variance in rates fluctuate according to topics and discourse factors; but unless a detailed conversational analysis is done, not much meaning can be attributed to these gaps in percentages.

In any case, the fact that the distributional pattern is so highly consistent within each generation, despite possible interference of discourse factors of all kinds is indisputable. With the exception of Hannah and Lida, we see very little variance in the preferred order of subject drop in terms of Person-Number within each generation. The result implies that even if the pattern found here could have started as some sort of individual style shifting, it has gained traction among speakers. This points to some specific community norms in action, thereby highlighting the fact that the Canberra Vietnamese community is a focused diaspora (§5.2.1.1), despite their modest number of speakers in comparison to other Vietnamese communities elsewhere (Chapter 2, §2.3).

### 5.7.1.3 Has Vietnamese co-evolved with speakers' English?

Given that we have seen signs of evolution of Canberra Vietnamese null subjects across generations, a final question to ask is whether this has evolved in tandem with speakers' English? Although we do not have benchmark data to confirm or refute contact effects (e.g. Nagy, 2015; Rinke et al., 2017; Torres Cacoulios & Travis, 2018), comparing the patterns between those directly in contact still allows us to gauge the extent to which these languages interact and influence each other.

<sup>107</sup> Although 'Interlocutor's Generation' did not emerge as a significant predictor in the multivariate analysis, this does not mean that this factor is not linguistically significant. As I discussed at the beginning of this chapter (§5.2.2), statistical modelling depends to an extent on the random distribution of the data and so statistical significance (or lack thereof) does not necessarily entail the presence or absence of linguistic meaningfulness (Kiesling, 2011, p.24).

A close look at the English dataset shows that the distribution of overt subjects in speakers' English is near categorical. In fact, out of more than 2,500 English clauses, there were only 39 instances of null subjects (Gen 1 = 25/1,380, Gen 2 = 14/1,202). All of these are either 2SG drop within an imperative clause (129), or within a conjoined clause with or without an overt conjunction (examples (130) and (131) respectively).

(129) Dany<sub>1</sub>: Wait a minute.

(Brian.Dany.0812, 02:45.8–02:50.4)

(130) a. Lida<sub>2</sub>: we just stopped talking,  
b. and then weren't friends anymore.

(Hannah.Lida.0718, 21:57.2–22:02.5)

(131) a. Reece<sub>1</sub>: we see a snake [...],  
b. catch them,  
c. cook them.

(Reece.Taylor.0906, 43:11.9–43:18.3)

Two facts can then be established: pronominal subjects are almost always expressed in speakers' English, and when a null subject occurs, it occurs in the expected environments that permit English null subjects (see Weir, 2012 for more discussion on this). This significantly differs from what we see in speakers' Vietnamese null subjects, both in terms of frequency and linguistic distribution. This observation is consistent with Nagy's conclusions for heritage Cantonese, Italian, and Russian in Toronto (2015), as well as Torres Cacoullos & Travis's (2018) recent work on New Mexican Spanish. The consensus is that the underlying grammar of subject drop in English and the substrate varieties remains separate, despite the highly bilingual nature of the communities and their sustained contact.

### 5.7.2 Beyond cross-generational variation: Stability of other conditioning factors

Having discussed the main result of cross-generational effects on null subjects, in this section I consider other conditioning factors that have emerged from Table 5.8. Here, space will be given to factors whose results cannot be straightforwardly understood, namely Coreferentiality (§5.7.2.1) and Primary language of the social network (§5.7.2.2), i.e. significant predictors for null objects. I will also draw attention to the robust distinction between null and overt forms across all variables (§5.7.2.3).

### 5.7.2.1 Coreferentiality effects

The first result that merits some discussion is the effect of Coreferentiality, which was selected as the strongest predictor for null objects. Specifically, null objects are favoured in cases where the antecedent is accessible from the preceding clause. Example (132) illustrates.

- (132) a. Luna<sub>1</sub>: *có gì Ø gọi me.*  
           have anything call 1SG.kin-MOTHER  
           ‘If anything (happens), then (you) call **me**.’
- b. Tressie<sub>2</sub>: *con gọi Ø,*  
           1SG.kin ask (2SG.kin-MOTHER)  
           ‘I called (**you**).’
- c.           *mà me có nói đâu.*  
           but 2SG.kin AFF say NEG  
           ‘But you didn’t say.’

(Luna.Tressie.0901, 3:21.9–03:23.7)

In this example, the object *me* ‘2SG-mother’ in (132b) is dropped in a coreferential context, referring to the same person in the preceding clause (132a). Given that Coreferentiality represents topicality, and topicality has been said to condition both subject drop (§5.5.2.1.3) and object drop (§5.5.2.2.1) in radical null subject languages, the fact that Coreferentiality is significant for objects but not for subjects is rather intriguing.

The first and seemingly most obvious explanation is a possible difference in data distribution and statistical modelling. Specifically, the modelling of subjects is more complicated (§5.5.1.1), with added factors of Interlocutor’s Age and Interlocutor’s Generation. These additions can inadvertently change the interactions between factors within the mix, making it more or less likely for any given predictor to rise above the significance line. This means that when all the relevant factors are considered, Coreferentiality is obfuscated as a condition for the realisation of Vietnamese pronominal subjects. In contrast, since the modelling for objects does not have these additional pragmatic factors, the effect of Coreferentiality is more readily observable.

Alternatively, this difference in results may actually reflect a linguistic difference between subjects and objects in radical pro-drop languages. In particular, the fact that Coreferentiality is the strongest predictor for Vietnamese null objects but is not selected **at all** for null subjects is in line with Huang’s (1984) proposal for Mandarin Chinese: null subjects and null objects differ in that null objects are more strongly bound to null topic in discourse. This point can be further explored in future work, where more fine-grained syntactic analyses can be carried out.

### 5.7.2.2 Different effects of environmental factors

The second factor that has emerged as significant for null objects is ‘Primary language of the social network.’ In particular, speakers who speak mainly Vietnamese within their social network are most likely to drop objects, followed by those with a mix of both languages. On the other hand, speakers who primarily use English within their social network disfavour Vietnamese null objects. Given that English allows null objects under stricter conditions than Vietnamese does, this finding suggests a role of frequency in input and usage in maintaining this variable. It also fits with the general consensus that cumulative and current exposure to the input of the heritage language significantly contribute to its grammatical outcome (Schmid, 2009; Unsworth, 2015).

That the primary language of the social network is only significant for null objects but not for subjects, however, is again illuminating. Similarly to what was explained for Coreferentiality, this might be partly due to the different modelling configurations for subjects and objects (§5.7.2.1). It is also conceivable, however, that there are other linguistic factors in play. For example, it may suggest that the frequency of usage and input have different levels of effects on different variables: some appear to be more sensitive to environmental factors than others under contact.

Cross-linguistically, this observation coincides with some evidence that the properties of [–null subject] are universally found in children, and tend to be earlier acquired than [ $\pm$ null object] (Wang, Lillo-Martin, Best & Levitt, 1992). Although the picture has emerged to be much more complicated than it seems (see Valian, 2016 for a helpful overview), this tendency generally holds and potentially shows a difference in timing of acquisition for the two variables. Linking this back to the findings in Chapter 4 (§4.6.3), where I showed how earlier acquired classifiers are better preserved under contact than their later acquired counterparts, results converge to support the prediction that the earlier a property is acquired, the more likely it is to remain present in speakers’ repertoire, independent of input. This possibly contributes to the explanation of why ‘Primary language of the social network’ plays a more significant role in maintaining Vietnamese null objects than null subjects.

### 5.7.2.3 A note on overt forms

Before concluding the chapter, I would like to return to the observation made in §5.6 about the distinction between null and overt forms in the corpus. Specifically, results in Table 5.4 are reproduced in Table 5.9 for illustration.

Looking at Table 5.9, there are two facts to establish. First, despite the differences in results for subjects, objects, and copulas, the distinction between null and overt forms remains robust across the board. Specifically, speakers from both generations produce both null and overt to-

Dependent Variable	Gen 1		Gen 2	
	N	%	N	%
<b>SUBJECTS</b>	4126	100%	818	100%
Null	1311	31.8%	258	31.5%
Overt	2815	<b>68.2%</b>	560	<b>68.5%</b>
<b>OBJECTS</b>	608	100%	384	100%
Null	145	23.8%	52	13.5%
Overt	463	<b>76.2%</b>	332	<b>86.5%</b>
<b>COPULAS</b>	671	100%	327	100%
Null	83	12.4%	31	9.5%
Overt	588	<b>87.6%</b>	296	<b>90.5%</b>

Table 5.9: Cross-generational distribution of null vs. overt subjects, objects, and copulas (Table 5.4 reproduced)

kens, suggesting that their knowledge of this variability remains intact. This pattern is applicable to all speakers in CanVEC, without any exception (overt subjects  $\bar{x} = 66$ ,  $s = 11$ ; overt objects  $\bar{x} = 77$ ,  $s = 8$ ; overt copulas  $\bar{x} = 87$ ,  $s = 5$ ).

Second, speakers' preference of overt forms over null forms is striking across all the three variables of interest. In particular, speakers produce overt subjects approximately 70% of the time, overt objects approximately 80% and overt copulas approximately 90% of the time. These numbers are significant, and even more so given that overt forms have been shown to exhibit distinctive behaviour in bilingual contexts of different kinds (see Sorace & Filiaci, 2006 et seq., Polinsky, 2018 and Aalberse et al., 2019 for a helpful overview). This means that although we now know quite well how null forms behave cross-generationally, the investigation remains incomplete until the overt forms are also examined.

## 5.8 Chapter summary

In this chapter, I used the variationist approach to investigate cross-generational variation in speakers' Vietnamese as a heritage language. Specifically, I probed three cases where the alternation of null and overt forms arises in Vietnamese: subjects, objects, and copulas. Results showed that while null objects and null copulas remained stable across generations, cross-generational variation was detected for null subjects. This contrasts with the initial impression given by raw rates, thereby highlighting the importance of looking beyond the surface patterns. More specifically, a closer investigation of linguistic and extra-linguistic conditioning factors reveals that the first-generation speakers were more likely to drop first-person subjects, while the second-

generation speakers were more likely to drop second-person subjects. Situating this in the community background, I explained this variation in terms of cultural integration into the Australian society. This evolution of subject pronominal use appears to have developed independently of patterns in speakers' English.

Beyond cross-generational variation, another notable finding was that speakers overwhelmingly preferred overt forms over null forms across all the three variables of interest. Given that overt forms account for at least around 70% of speakers' production, and that it has frequently been suggested that the overt counterparts of null forms exhibit distinctive behaviour in bilingual contexts of different kinds, it is imperative that we explore them in more detail. This is hence the focus of the next chapter.



---

## PROBING INTERFACE VULNERABILITY: ON THE (OVER)USE OF OVERT FORMS

---

### 6.1 Introduction

Having examined the cross-generational patterns of **null** forms of subjects, objects and copulas in the previous chapter, I now turn to the cross-generational patterns of **overt** forms. Key attention will be given to the interface components, i.e. components that link different sub-modules of language, or language and other non-linguistic cognitive systems (Chomsky, 1995, p.2).<sup>108</sup> These interface components have been shown to exhibit distinctive behaviour in various scenarios of language contact (e.g. Sorace & Filiaci, 2006 et seq.). While the variationist approach adopted in [Chapter 5](#) may allow researchers to probe the **probability** and **patterns** of speakers' use of overt forms, in this chapter, I appeal to the interface-oriented approach to offer **a more fine-grained insight** into the extent to which overt forms are used in a heritage-language context. Specifically, the interface-oriented approach brings the focus back to the underlying cross-linguistic factors that potentially condition the vulnerability of different phenomena under contact. Given that the interface-oriented approach has also featured strongly in recent generative discussion of the increased use of overt forms in relevant communities, what I primarily aim to achieve in this chapter is to identify whether the Canberra Vietnamese bilingual community also exhibits interface vulnerability effects of the kind that have been uncovered in other bilingual landscapes.

---

<sup>108</sup>Different sub-modules of language include phonology, lexicon, morphology, syntax, semantics, pragmatics, etc., and non-linguistic cognitive systems include components such as systems of vision or motor planning, etc.

The discussion begins with [Section 6.2](#) which provides background on the behaviour of overt forms in heritage languages. [Section 6.3](#) follows with a description of the coding procedure. [Section 6.4](#) presents the results, before [Section 6.5](#) concludes the chapter.

## 6.2 Background

In this section, I discuss previous findings on the patterns of usage of overt forms (as opposed to the corresponding null forms) in heritage languages. Specifically, I focus on two main areas which recent work has elaborated on:

- (i) the over-extension of the pragmatic contexts of overt forms (§6.2.1); and
- (ii) the interface components that underlie vulnerability under contact (§6.2.2).

Before we proceed, however, it is important to remember that in the present work, both Gen 1 and Gen 2 speakers are considered Vietnamese heritage language speakers, and both their varieties are considered Vietnamese heritage language for all intents and purposes (see [Chapter 1](#), §1.1 for my motivation).

### 6.2.1 The over-extension of pragmatic contexts of overt forms

Various studies have reported an increase in overt forms of bilinguals of all kinds, especially in contexts where a null pronominal form would be more appropriate (e.g. Silva-Corvalán, 1994; Sorace, 2000; Müller & Hulk, 2001; Montrul, 2002; Tsimpli, Sorace, Heycock & Filiaci, 2003; Montrul, 2004; Sorace, 2004; Tsimpli, Sorace, Heycock & Filiaci, 2004; Sorace & Filiaci, 2006; Otheguy et al., 2007; Rothman, 2007; Montrul, 2008; Rothman, 2009; Sorace, 2011; Otheguy & Zentella, 2012; Ivanova-Sullivan, 2014; Montrul, 2015; Quesada, 2015; Polinsky, 2018). Sorace (2000), for example, examined the pragmatic distribution of null and overt subject pronominal forms in Italian near-native English speakers (i.e. an equivalent of Gen 1 baseline speakers in this study) and found that bilingual and monolingual speakers behave differently in relation to overt forms. Specifically, her results show that when presented with a leading question such as in (133a), monolingual speakers residing in Italy tend to produce structures with a null form as in (133b), while bilinguals often produce utterances with an overt form as in (133c).

- (133) a. *Perchè Maria è uscita?* [Leading Q]  
           why Maria is left  
           ‘Why did Maria leave?’

- b.  $\emptyset$  *ha deciso di fare una passeggiata.* [Monolinguals]  
 has decided to do a walk  
 ‘(She=Maria) decided to go for a walk.’

- c. *Lei ha deciso di fare una passeggiata.* [Bilinguals]  
 She has decided to do a walk  
 ‘She (=Maria) decided to go for a walk.’

(Examples reproduced from Sorace, 2000)

This means that the increase of overt forms begins already in the first-generation immigrants, whose language serves as input for second-generation heritage language speakers (e.g. Otheguy et al., 2007; Dubinina & Polinsky, 2013; Polinsky & Scontras, 2020). According to Polinsky (2018, p.7), these incipient changes in the input are amplified in heritage language speakers, as reported for Gen 2 speakers of other heritage varieties of numerous languages such as Turkish (Backus, 2005), Hungarian (Bolonyai, 2000), Greek (Lozano, 2006; Margaza & Bel, 2006), Mexican Spanish (Montrul, 2004), and Faetar (a Francoprovençal dialect in southern Italy; Nagy, Iannozzi & Heap, 2017). The general consensus is that while the syntactic availability of null subjects is acquirable and retainable<sup>109</sup>, the conventions underlying the interpretation and production of overt subjects are substantially blurred in heritage language varieties (e.g. Mohring & Meisel, 2003; Tsimpli et al., 2004; Montrul, 2005, 2006; Müller, 2007; Perez-Cortes, 2018; Lustres, 2018).

This observation of overt form overuse has recently been characterised as the ‘Silent problem’ or the ‘Avoidance of Ambiguity,’ particularly for second-generation bilinguals (Polinsky, 2018; Polinsky & Scontras, 2020). In particular, studies have shown that second-generation heritage language speakers prefer categorical one-to-one mappings between form and function, and when the form is ambiguous or unrealised altogether, this creates interpretational problems. Laleko & Polinsky (2016, 2017), for example, show that Japanese heritage language speakers appear native-like in their production and comprehension of the Japanese *-wa* for contrastive topic, but struggle when it comes to identifying the proper function of the same marking for thematic topics. They attribute this difference in linguistic outcomes to the difference in the scope of ambiguity between these two types of topic function. Examples (134)–(136) demonstrate the distinction.

- (134) a. Context: A family moved into the apartment next to mine. They have a 10-year-old girl and a 6-year-old boy. The girl usually stays inside and rarely comes out, and I have never heard her talk.
- b. *Otoko-no ko-wa totemo genki-da.* [Contrastive]  
 man-GEN child-TOP very healthy-COP.nPST  
 ‘THE BOY is very active.’

<sup>109</sup>In fact, this has already been reflected in CanVEC: As we see in Chapter 5, the availability of null subjects, objects, and copulas remains robust across generations: ~30% of null subjects, 13.5%–23.8% for null objects, and 9.5%–12.4% for null copulas.

(135) a. Context: A family moved into the apartment next to mine. They have two boys, a 10-year-old and a six-year-old. They are always running around the apartment complex, doing all sorts of things. Whenever I hear them chasing each other outside of the apartment, I say to myself:

b. *Otoko-no ko-wa totemo genki-da.* [Thematic-GENERIC]  
 man-GEN child-TOP very healthy-COP.nPST  
 ‘Boys are very active.’

(136) a. Context: A family moved into the apartment next to mine. They have two children.

b. *Otoko-no ko-wa totemo genki-da.* [Thematic-ANAPHORIC]  
 man-GEN child-TOP very healthy-COP.nPST  
 ‘The(ir) boy is very active.’

(Examples reproduced as per the original from Polinsky & Scontras, 2020, pp.9–10)

According to Polinsky & Scontras (2020), contrastive and thematic markers differ in their scope of ambiguity. Specifically, contrastive topics as in (134) are restricted to the negation of at least one of the alternatives (and are hence more categorical), while thematic topics are more variable: they can be either GENERIC (i.e. referring to a class of entities not explicitly linked to prior discourse, as in (135)) or ANAPHORIC (i.e. referring to entities previously mentioned in the discourse, as in (136)). This one-to-many mapping between form and meaning in the thematic scope is expected to create difficulties for heritage language speakers. Linking this back to the case of pronominal use, we can draw a parallel in that there too exists a more categorical variant and a more variable variant. In Vietnamese, for example, overt forms encode more concrete information that facilitates one-to-one mappings (like the contrastive *-wa*), while null forms leave information unexpressed and therefore involve more ambiguity (like the thematic *-wa*). Consider the following example:

(137) a. *lúc em<sub>1</sub> điện cho anh<sub>2</sub> là em<sub>1</sub> qua tới gần đường-cao-tốc rồi.*  
 when 1SG.kin call for 2SG.kin COMP 1SG.kin past to near high-way PERF  
 ‘When I called **you**, I had already gone past the highway.’

(Brunelle’s data, 2020, glosses and translations mine)

b. *lúc em<sub>1</sub> điện cho anh<sub>2</sub> là Ø<sub>1/2/x</sub> qua tới gần đường-cao-tốc rồi.*  
 when 1SG.kin call for 2SG.kin COMP past to near high-way PERF  
 ‘When I called **you**, (?) had already gone past the highway.’

In (137a), the overt pronominal form *em* in the main clause makes it clear who had gone past the highway when the phone call was made. In contrast, when this pronominal form is left unrealised as in (137b), ambiguity occurs: the empty subject can refer back to either *em* or *anh* in the subordinate clause, or even to somebody else previously mentioned in discourse.<sup>110</sup> In this

<sup>110</sup>Thank you to Linh Hoàng, Phi Hoàng and Mai Nguyễn for their native judgements of this utterance.

sense, overt pronouns facilitate one-to-one mapping between the anaphor and the antecedent, while null pronouns may produce one-to-many mappings.<sup>111</sup> Previous research has established that, as a general tendency, heritage language speakers have been found to avoid this ambiguity by reducing the use of ambiguity-triggering elements (such as Vietnamese null pronominal forms or Japanese thematic *-wa*) to a more restricted role in discourse, or even do away with them altogether in extreme cases (Laleko & Polinsky, 2017; Polinsky, 2018; de Prada Pérez, 2019; Polinsky & Scontras, 2020).

For copulas, although ambiguity is different from pronominal-form ambiguity, researchers have also observed an extension of use of one form over another in some heritage varieties (e.g. Silva-Corvalán, 1994; Gutierrez, 2003; Salazar, 2007; Carter & Wolford, 2018). Spanish copulas are a good case in point. Specifically, there are two forms of copula in Spanish: *ser* and *estar*, the former of which is used to express an inherent quality that remains permanent, while the latter refers to situational qualities that are subjective to the speaker (e.g. ‘*Es feliz*’ for s/he is happy by nature vs. ‘*Está feliz*’ for s/he is happy now). Various studies, however, have documented a ‘change in progress,’ where younger speakers were found to consistently use *estar* in contexts where *ser* is expected. Particularly relevant to this discussion is the work of Gutierrez (2003), which compares the use of *ser* and *estar* across different generations in Los Angeles (Silva-Corvalán, 1994) and Houston (data Gutierrez personally obtained from Alejandra Balestra and Jennifer Ayres) against monolingual speakers from Morelia, Mexico. Gutierrez’s results show that although the extension of contexts where *estar* can be used originates in the monolingual community, it advances at a faster rate in the bilingual communities (16% of innovative *estar* in Morelia (N=139/866) vs. 22% in Houston and 34% in Los Angeles), particular among younger generations.<sup>112</sup> He further shows that while innovative *estar* is only restricted to certain environments such as description, age, size, physical appearance, and evaluative adjectives in monolingual communities, it has extended to also cover moral characteristics, social status, perception, and colour in bilingual communities. In other words, this extension of use appears to be in a more advanced state in the heritage language. This finding has since been corroborated by other

<sup>111</sup>Note, however, that the phenomenon is complex and there are of course cases where overt forms can be ambiguous too. An example of this is when the pronominal form indexes features that can be bound to multiple antecedents; e.g. ‘John called Mark when he finished work,’ where ‘he’ could refer to either John or Mark in Vietnamese and English. In Vietnamese speech, however, speakers rarely use the Vietnamese 3SG pronoun *anh* ‘he’ in such circumstances, but rather prefer explicit personal names; e.g. ‘John called Mark when John/Mark finished work. When the subject of the main clause is left null, the sentence again becomes ambiguous as in the case of (137b); i.e. it could be Mark, or John, or somebody else who finished work.

<sup>112</sup>Sample size (in terms of tokens) for Houston and Los Angeles is not reported in Gutierrez’s (2003) study. There also seems to be an oversight in the citation of the rates of innovative *estar* in Los Angeles, as this figure is 55% (N=344/623) in the original article by Silva-Corvalán (1986). This, however, does not affect Gutierrez’s argument.

studies of young bilingual heritage Spanish speakers in the U.S, whom are found to show the highest use of innovative *estar* (Salazar, 2007; Carter & Wolford, 2018).<sup>113</sup>

In relation to Vietnamese, although there is no dedicated habitual copula as such,<sup>114</sup> the realisation versus non-realisation of copula *là* in an AdjP environment similarly triggers an interpretative distinction. Specifically, as set out in Chapter 5, §5.3.3, while null copulas are preferred for an AdjP, overt forms are used for emphasis. Examples (93) and (94) are repeated below in (138) and (139) respectively to illustrate.

(138) *Gói hàng này là rất nặng*  
 CLS goods DEM COP INTSF heavy  
 ‘This package (of goods) is very heavy.’

(139) *Gói hàng này là nặng rồi*  
 CLS goods DEM COP heavy PERF  
 ‘This package (of goods) is already heavy.’

In (138), the COP *là* is selected together with the intensifier *rất* ‘very.’ Similarly, the perfective particle *rồi* ‘already’ also co-occurs with *là* in (139). Omission of the intensifier or the perfective particle in these cases renders the sentence unacceptable (Chapter 5, §5.3.3.)

Given that:

- (i) the use of overt forms is often inflated in contact situations generally; and
- (ii) the pragmatic context of one form is often extended to cover (parts of) the other,

we might expect that Vietnamese speakers in Canberra, especially second-generation speakers, extend the use of the copula *là* in AdjP to cases where emphasis is not marked (i.e. where an appropriate particle and/or intensifier is not deployed).

## 6.2.2 The vulnerable nature of the interfaces

The observation that pragmatic components are often affected in the heritage language has been taken to suggest that different language modules are subject to different levels of ‘vulnerability’ under contact (e.g. Sorace & Filiaci, 2006; Sorace & Serratrice, 2009a; Sorace et al., 2009b; Sorace, 2011; Tsimpli, 2014; Sorace, 2016). For early-bilingual heritage language speakers in particular, it has been observed that they tend to retain ‘the basic, perhaps universal, core struc-

<sup>113</sup>The fact that it is *estar*, i.e. the option with fewer specified features, which is over-generalised to domains where *ser*, i.e. the option with more specified features, is required suggests a direction of over-generalisation that is consistent with what is proposed by the Maximise Minimal Means (MMM) model (Biberauer, 2017, 2018, 2019). Specifically, the MMM model highlights speakers’ acquisition and grammar-shaping bias to maximally utilise minimal resources. This is taken as a general cognitive bias that is active not just in child acquirers, but also in adults.

<sup>114</sup>See Clark (1996), however, for a few examples where the topic marker *thì* contrasts with the copula *là* in a similar way to *ser* and *estar*.

tural properties of their language,' while showing vulnerability in other domains (Benmamoun, Montrul & Polinsky, 2013, p.148). Although it is not always clear what the basic, universal and core structures of language are, core grammatical properties have often been identified with syntax, while non-core properties involve interfaces with different language modules as well as with extra-linguistic considerations (Tsimpli, 2014, p.286). In this context, 'interface' refers to a component that links different sub-modules of language, or language and other non-linguistic cognitive systems (Chomsky, 1995, p.2). [Figure 6.1](#) illustrates the domain where core and non-core components belong in a generative model of language architecture.

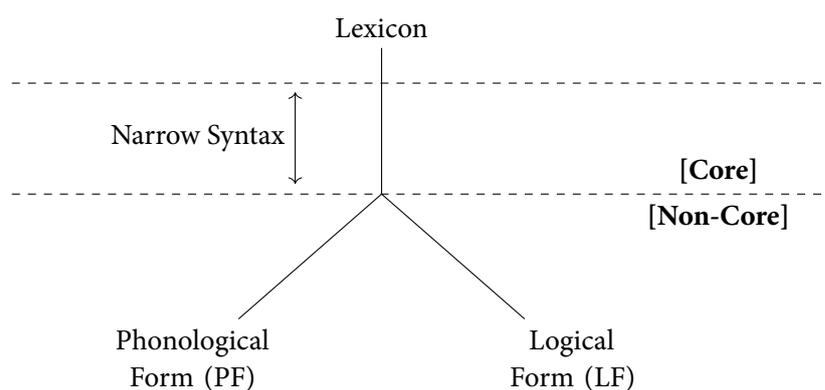


Figure 6.1: Core and non-core components in the Y-model of architecture of language

As [Figure 6.1](#) demonstrates, core components operate within narrow syntax, i.e. before any syntax-external computation (e.g. at PF or LF) occurs, while non-core components relate to what happens post-syntactically, particularly along the LF path.<sup>115</sup> In this sense, what belongs to narrow syntax is considered core, while most of what belongs outside is non-core and therefore more vulnerable. Works by Tsimpli & Sorace (2006), Sorace et al. (2009b), and Serratrice, Sorace, Filiaci & Baldo (2009) further differentiate between **internal** and **external** interfaces. For instance, they write:

The distinction between the two interfaces is based on the assumption that the syntax-discourse interface is a 'higher' level of language use, integrating properties of language and pragmatic processing, whereas syntax- semantics involves formal properties of the language system alone.

(Tsimpli & Sorace, 2006, p.653)

In this sense, the internal interface denotes a component that combines syntax and semantics, while the external interface denotes a component that relates syntax and discourse-pragmatics. Tsimpli & Sorace (2006) further proposed that while violation at the external interface is typi-

<sup>115</sup>Note that in [Figure 6.1](#), lexicon lies outside the formal system under discussion here.

cally a matter of gradient acceptability (e.g. the alternation between null and overt pronouns in pro-drop languages), some violations of internal interface requirements can result in clear ungrammaticality (e.g. Focusing in Greek, which is syntactically encoded and involves an operator-variable dependency).

From a formal perspective, this difference is a result of the difference in the predictability of the required mapping: while the mapping between syntax and semantics is relatively fixed, the mapping between syntax and pragmatics is open to more possibilities. From a functional perspective, this difference can be attributed to processing costs: the more information from different modules that a property requires, the more difficult it is to acquire and process. Although the starting point is somewhat different for Sorace & Filiaci's interface proposal and Tsimpili's distinction between core and non-core properties, the key idea is common to these two accounts: syntax is core and most stable, while larger interface domains are non-core and most susceptible to change. This is consistent with the general observation in our previous discussion that early-bilingual heritage language speakers appear native-like when it comes to categorical one-to-one mapping (§6.2.1), while deviating from monolingual speakers when the usage involves more nuances.<sup>116</sup>

In the context of this work, it is clear that the overt forms of Vietnamese subjects, objects, and copulas all require additional resources beyond narrow syntax, albeit to different degrees. Specifically, as we have seen in Chapter 5 (§5.3.1), the production of overt Vietnamese pronominal forms requires not only consideration of discourse-pragmatic components of coreferentiality, but also language-external knowledge such as judgement of the referent's age and social status relative to the speaker. This sort of knowledge is relatively complex and requires social judgement and the integration of information of different kinds. Overt copulas, on the other hand, do not demand this level of resource and might therefore be expected to remain stable. Linking this back to the distinction between internal and external interfaces, Table 6.1 captures the nature of each non-core component required to regulate the realisation of overt subjects, objects, and copulas in Vietnamese.

It is clear from Table 6.1 that all the three variables of interest encode extra-syntactic components to some degree. Overt copulas in particular only clearly require some mapping at an

<sup>116</sup>The key idea here is also apparent in several other hypotheses in the field, such as the Transparency Hypothesis (O'Grady, Kwak, Lee & Lee, 2011), the Vulnerability Hypothesis (de Prada Pérez & Pascual y Cabo, 2012), and the 'Smaller Domain Principle' (Koornneef, Avrutin, Wijnen & Reuland, 2011; Reuland, 2011). The Smaller Domain Principle, for example, posits an implicational hierarchy:

Narrow Syntax < Logical syntax (Conceptual-Intentional interface) < Discourse

In terms of this hierarchy, linguistic components furthest to the left involve least mapping, and therefore require fewer resources to acquire, process, and retain, while components further to the right incrementally involve more mapping and therefore more resources to remain stable.

Variable	Component		
	Internal interface	External interface	
	Syntax-semantics	Syntax-discourse	Syntax-pragmatics
Overt subjects	–	Coreferentiality	Referent's age and gender
Overt objects	–	Coreferentiality	Referent's age and gender
Overt copulas	Emphasis	–	–

Table 6.1: Interface association of overt subjects, objects, and copulas in heritage Vietnamese. (–) denotes non-applicability. The cell colours indicate degrees of vulnerability: the darker the colour, the more vulnerable the component is.

internal interface level, specifically emphasis marking when it is realised with an AdjP predicate. As previously described (Chapter 5, §5.3.3), the realisation of copulas in an AdjP environment requires the co-occurrence of an intensifier or perfective particle in order to achieve emphasis. When the intensifier or perfective particle is not realised in these environments, the sentences are not really considered acceptable (Diep, 2004, p.103). The overt copula and additional particle pairing thus represents a relatively fixed mapping, and given that the violation at this level can give rise to clear ungrammaticality, this component represents an internal interface by definition (Tsimpli & Sorace, 2006). In the domain of interfaces, this is the area that is expected to remain most stable.<sup>117</sup>

In contrast, Vietnamese overt subjects and objects require resources associated with the external interfaces. Here I make a finer distinction between **discourse** and **pragmatics**: **discourse** refers to an organised set of utterances, which together create textual production of a particular meaning, whereas **pragmatics** refers to a particular use in human communicative practice, i.e. a performance that is socially and culturally regulated (Blommaert, 2011, pp.126-127). In this sense, **discourse** features in this study are non-core features that reside within a coherent set of actual instances of language use (such as coreferentiality), and **pragmatics** features are those that require implicit knowledge completely external to linguistics (such as knowledge of social norms and of the interlocutors). The syntax-pragmatic external interface in particular is inherently complex, and is thus also where we expect the most vulnerability.

The Coreferentiality component of Vietnamese pronominal subjects and objects (both null and overt forms) requires mapping to the syntax-discourse external interface, and is also expected to be vulnerable (Sorace, 2011; Papadopoulou, Peristeri, Plemenou, Marinis & Tsimpli, 2015). This is because speakers have to simultaneously assess shared knowledge with interlocutors, integrate contextual information, update the mental representation of the situation and use

<sup>117</sup>It should be noted that while we probably cannot firmly exclude the relevance of discourse factors in relation to copulas, the contribution of this element to the realisation of copulas in Vietnamese is still little discussed. More research in this domain is needed, but this lies beyond the scope of the present discussion.

this information to configure a precise form to meaning mapping. In the context of bilinguals, we might expect that the processing demand is even greater, if the two languages are always actively competing for resources in a bilingual's mind (Green, 1998; Costa, Caramazza & Sebastian Galles, 2000; Dijkstra & van Heuven, 2002; Marian & Spivey, 2003).<sup>118</sup> Furthermore, given that bilingual speakers have to split their time between two languages, their input quantity is considerably reduced in comparison to that of monolingual speakers (Unsworth, 2013; Polinsky & Scontras, 2020). In the context of a heritage language (which is often a minority language, such as Vietnamese in Canberra), the second-generation heritage language speaker's input in the heritage language is also limited to a finite set of speakers and topics for which the heritage language is commonly used. According to Tsimpli (2014, p.284), although half of the 'normal' amount of input (i.e. the amount of input that a monolingual typically receives) is sufficient for the core, early-acquired elements of the language (such as the availability of null subjects, [Chapter 5](#)), this reduced environmental support will have an effect on the acquisition and retention of the interface, late-acquired phenomena such as pronoun resolution.

### 6.2.3 Summary

Overall, this section has highlighted two basic facts about overt forms in heritage language. First, the pragmatically conditioned distribution between null and overt forms can become substantially blurred for speakers of a heritage language (both Gen 1 and Gen 2 speakers), with the distribution of overt forms often extending to contexts where null forms are expected. Second, the further a component is removed from core syntax (i.e. the early-acquired component), the more vulnerable it appears to be.

In relation to heritage Vietnamese, we thus expect to see some variation across all the three variables of interest. Phenomena we might expect to see include:

- (i) overt pronominal forms being used in contexts where null forms are typically preferred in monolingual Vietnamese;
- (ii) inappropriate overt pronominal types (i.e. pronominal forms, kin terms, or names) being used, e.g. neutral pronominal forms in place of kin terms, or vice versa;

---

<sup>118</sup>In a study on Spanish, for example, Otheguy & Zentella (2012) showed that while 6–12 year old monolingual children in Mexico tend to overuse null subjects in switch reference contexts, they no longer do so when they are about 13 to 14 years old and instead use null and overt subjects in a manner similar to adults. This shows that monolingual children also experience difficulties with the discourse properties of null and overt subjects, and so if they already take that long to converge on adult grammar in a monolingual context, this can only be expected to be more problematic in a bilingual environment.

- (iii) inappropriate overt pronominal forms (i.e. forms that inaccurately index the gender/age of the referent) being used; and
- (iv) overt copulas being used with an AdjP predicate where emphasis is not marked.

In order to test these predictions, I next explicate how overt subjects, objects, and copulas are analysed.

## 6.3 Analysing CanVEC overt forms

### 6.3.1 Defining appropriateness

The focus of this chapter is to explore the cross-generational variation concerning different interface factors regulating the occurrence of overt subjects, objects and copulas in Vietnamese. In doing so, I consider whether each token is used appropriately in a given context. In this section, I discuss how ‘appropriateness’ is broadly defined, before describing in detail the coding procedure for each of the variables of interest.

For overt subjects, objects, and copulas, ‘appropriateness’ is first assessed on the basis of whether the overt form is redundant, i.e. whether the use of null forms is natural in a given context. As I will show in §6.3.2.1 and §6.3.3.1, this is not always straightforward for Vietnamese pronominal forms. For overt subjects and objects in particular, ‘appropriateness’ is also further defined, given the complexity of the information indexed in the choice of pronoun types and pronoun forms in discourse. I have discussed the nuances of these choices in various places in [Chapter 3](#) and [Chapter 5](#). Here, I elaborate and summarise these pragmatic distinctions in [Table 6.2](#).

As [Table 6.2](#) demonstrates, each pronominal type in Vietnamese carries a specific pragmatic load in terms of contextual appropriateness. This is further complicated by the rich indexicality of age and gender in kin terms and certain forms of pronouns particularly. The elaborate nature of the considerations underlying the choice of pronominal forms in Vietnamese is captured in [Figure 6.2](#). Recall from [Chapter 5](#), §5.5.1.1 that in this work, I make no distinction between referents and participants. The term ‘referent’ is used independently of grammatical person, and can cover self- (1SG), interlocutor- (2SG) and third-party- (3SG) reference.

Pragmatic condition	Pronouns	Kin terms	Personal names
Respect	×	✓	–
Family setting	×	✓	*
Younger/similar-aged interlocutor/3SG referent	✓	✓	✓
Older interlocutor/3SG referent	×	✓	×
Emotional distance	✓	*	×
Casual/friendly	*	✓	✓

Table 6.2: Pragmatic distinctions between different pronominal types in colloquial Vietnamese. Check mark (✓) indicates desired effects/appropriate use, cross mark (×) indicates opposite effects/inappropriate use, dash (–) indicates non-applicability, and asterisk (\*) indicates possible usage in restricted contexts.

As Table 6.2 and Figure 6.2 together illustrate, the choice of certain pronominal types and pronominal forms involves multiple levels of information. Appropriateness for the overt use of pronominal subjects and objects is thus further defined by two additional criteria:

1. Does the chosen pronominal type (i.e. kin term, pronoun, personal name) occur in the expected environment? and
2. Does the chosen pronominal form index the correct information (i.e. the correct age cue and gender of the referent)?

For each variable, examples will be given for cases that are deemed appropriate and inappropriate use. Given my background as a contact speaker, all my judgements were cross-validated by non-contact speakers to ensure validity. Unless otherwise stated, all the coding in this chapter is entirely manual.<sup>119</sup>

## 6.3.2 Coding overt pronominal subjects

### 6.3.2.1 Redundant overt pronominal subjects

For overt subjects, I first consider cases of overt pronominal forms where null forms are typically expected. In consistent null subject languages such as Spanish and Italian, research has

<sup>119</sup>Specifically, 20% of each variable was independently coded by another native Vietnamese speaker living in Vietnam. After that, results were compared with mine for inter-rater reliability. The agreement rates range between 93% and 97% for each dataset. Where there was a mismatch, it was taken to another native speaker, and the final judgement was the one given by the majority. Thank you to Luong Xuan Vu, Linh Tuyen Hoang, and Phi Hoang for their native judgements on different datasets. These informants were only given access to the anonymised data with no information about the speakers.

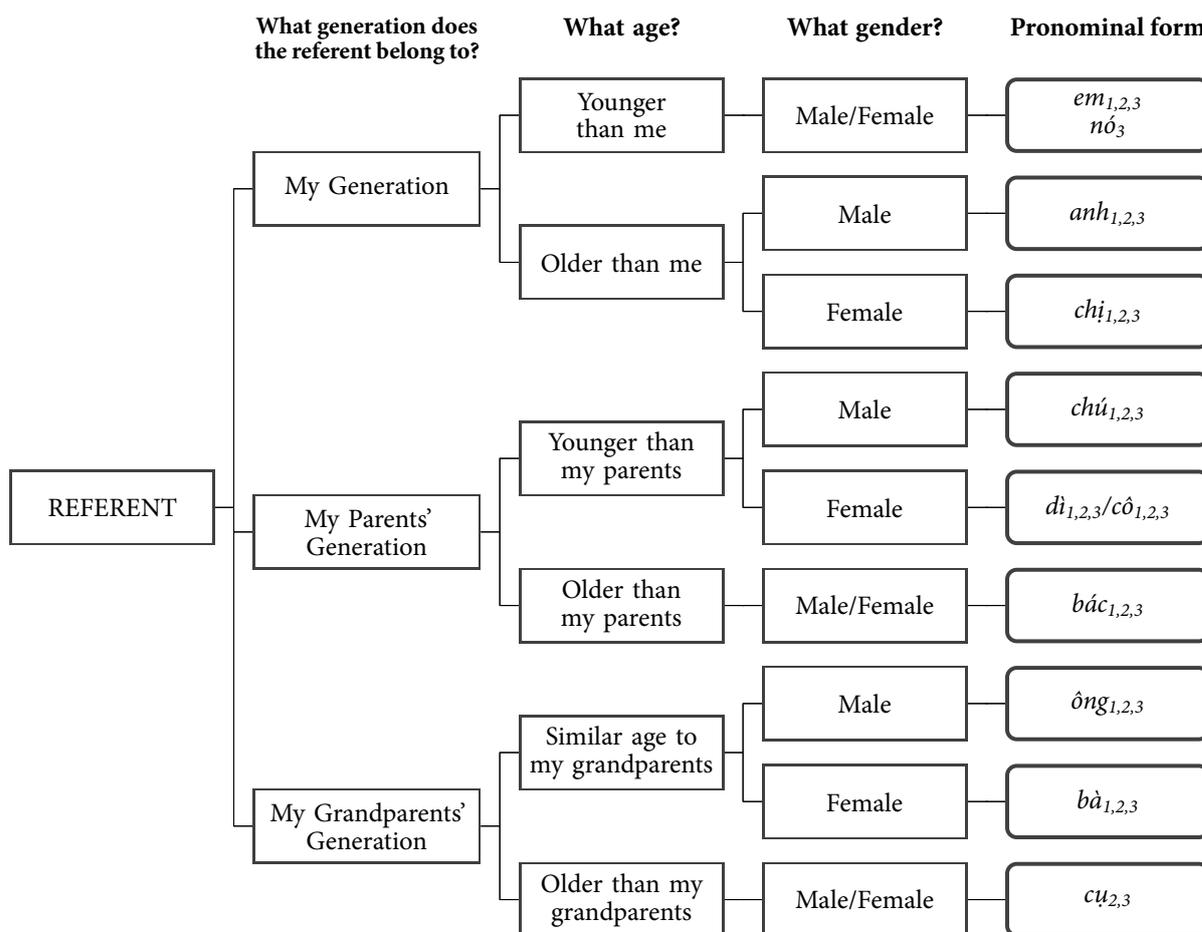


Figure 6.2: A simplified version of a gender and age index for some basic Vietnamese pronominal forms. The subscripted number denotes the grammatical person that each form can index. For example, *em* can be first-, second-, or third-person depending whom it refers to in discourse, whereas *nó* is exclusively third-person (see also §5.3.1 and §5.4.2).

established a (relatively) clear division of labour between null and overt pronouns. Specifically, null pronouns are typically believed to be bound to referents in subject position, while overt pronouns are used for object referents and subjects that receive emphasis for some reason (e.g. contrast) (see e.g. Alonso-Ovalle, Fernández-Solera, Frazier & Charles Clifton, 2002; Carminati, 2002). In a radical pro-drop language such as Vietnamese, however, the distinction is not always equally straightforward. In fact, several studies have stated that the pro-drop in radical null subject languages is only partially understood (Jia & Bayley, 2002; Ngo, 2019), and that the division of labour between null and overt pronouns in Vietnamese is less clear than sometimes assumed (Ngo, 2019). In the context of CanVEC, the most clear-cut case where the ‘redundancy’ of overt subject pronominal forms is well-attested is where **both** of the following conditions are satisfied:

- (i) the referent remains unchanged (i.e. where there is no need for disambiguation between multiple possible referents or reintroduction of a referent); AND

(ii) there exists some shared structure and/or main verb with the preceding discourse.

Consider example (140) as a case in point.

- (140) a. Tressie<sub>2</sub>: Lala *hơi khó hơn* with people. [Context]  
           a-bit difficult more  
           ‘Lala is little more difficult with people.’  
           [10]
- b. Luna<sub>1</sub>: *nó nói tiếng Anh đó,* [Appropriate]  
           3SG speak language England DM  
           ‘She speaks English,’
- c. *nó nói dữ lắm.* [Inappropriate]  
           3SG speak much INTSF  
           ‘She speaks a lot.’

(Luna.Tressie.0901, 09:36.4–09:44.6)

As we can see, in this example the referent of *nó* remains unchanged (=Lala), while the main verb *nói* ‘speak’ in (140b) is repeated in (140c) in a similar construction. In this context, the repetition of the subject pronominal form *nó* in (140c) is thus unduly emphatic, and hence deemed inappropriate. Overt subject pronominal forms in all other cases (including instances where the VPs/general structures are not shared) are otherwise marked as appropriate. This is obviously a rather restrictive definition of ‘inappropriate’ for overtly realised subject pronouns, which might mean that inappropriate overt subject use may be under-reported in this work. This limitation, however, is inevitable given the current state of our (lack of) knowledge on when precisely it is and is not acceptable to use overt pronouns in a radical pro-drop language such as Vietnamese (Jia & Bayley, 2002; Ngo, 2019). Establishing well-defined criteria to evaluate this phenomenon is the natural next step, but that lies beyond the scope of the present dissertation.

### 6.3.2.2 Type of pronominal subjects

The second element to be coded for is the type of subject pronominal form. Table 6.3 gives an overview of the distribution of different types of overt subjects in CanVEC.

Overt subjects	Gen 1		Gen 2	
	N	%	N	%
<b>Pronouns</b>	359	13%	68	12%
<b>Kin terms</b>	1,666	59%	455	81%
<b>Personal names</b>	797	28%	37	7%
<b>TOTAL</b>	2,822	100%	560	100%

Table 6.3: Distribution of different pronominal types for overt subjects in CanVEC

As Table 6.3 demonstrates, kin terms are consistently the most frequent type across both generations. Pronouns and personal names, nonetheless, are also productively used in the corpus. The purpose of looking more closely at overt subject-type is to consider whether the choice of a certain pronominal type is appropriate in a given context. Example (141) provides an illustration.

- (141) a. Brian<sub>1</sub>: **Dany** *đừng đánh-giá người nào cũng qua chuyện học hết.* [Inappropriate]  
 Dany NEG judge person any also through CLS study DM  
 ‘You-DANY shouldn’t judge anyone by their academic achievement.’  
 (lit. ‘You shouldn’t judge anyone through their studies.’)

- b. Dany<sub>1</sub>: **chị Dany** *đâu có đánh-giá đâu.* [Appropriate]  
 1SG.kin Dany NEG AFF judge DM  
 ‘I-SISTER-DANY didn’t judge.’

(Brian.Dany.0812, 17:52.4–17:59.3)

In this example, Brian is Dany’s younger brother, and is expected to refer to Dany using a kin term. However, he instead uses a proper name to refer to Dany (141a), and so the overt subject pronominal form is marked as inappropriate. In contrast, Dany responded using a combination of kin term and personal name (*chị Dany* ‘sister Dany’), and so her utterance is marked as appropriate.

It is important to note, however, that there are cases where the selected pronominal type might seem ‘inappropriate,’ but is in fact deployed in an appropriate context. Example (142c) illustrates.

- (142) a. Luna<sub>1</sub>: **me** *nói con hoài.* [Appropriate]  
 1SG.kin say 2SG.kin often  
 ‘I-MOTHER have said it to you-CHILD a lot.’

- b. **mà con** *không nghe.* [Appropriate]  
 but 2SG.kin NEG listen  
 ‘But you-CHILD didn’t listen.’

- c. **chắc tui** *buồn chết.* [Appropriate]  
 maybe 1SG sad die  
 ‘I might be sad to death.’

(Luna.Tressie.0901, 07:15.5–07:20.9)

In this example, Luna is talking to her child, Tressie. Although Luna used kin terms to refer to herself and her child in (142a) and (142b) as expected, she changed the type of subject expression to a personal pronoun *tui* ‘1SG’ in (142c). While this is typically not expected in a family setting (and hence seems inappropriate), the switch of pronominal form type here is employed for a specific stylistic effect. In particular, given that Tressie did not listen to Luna, Luna’s switch to the personal pronoun here has the effect of distancing herself from her daughter. As such, the use of the first-person *tui* here does not indicate Tressie’s compromised competence, but indeed

her sensitivity to the complexity of this system. Cases like this (N=5) are therefore all counted as cases of appropriate use.

### 6.3.2.3 Form of pronominal subjects

In addition to pronominal type, I also marked certain pronominal forms for appropriateness in relation to gender and age (N=2,333).<sup>120</sup> As we previously saw in Figure 6.2, each term carries information about the perceived age range and gender of a referent. These pronominal forms are thus marked for whether they index the correct gender and age of the referent. Consider example (143).

- (143) a. Max<sub>1</sub>: *mà sao anh phải chuyển qua đây?* [Appropriate]  
 but why 2SG.kin must move across here  
 ‘But why did you-MALE.OLDER have to move here?’
- b. Thomas<sub>1</sub>: *nói chung là bên Mỹ anh không thích.* [Appropriate]  
 speak general COMP side America 1SG.kin NEG like  
 ‘Generally speaking, I-MALE.OLDER didn’t like America.’  
 (Max.Thomas.0823, 01:04.8–01:29.5)

In this example, Thomas is two years older than Max (Chapter 2, Table 2.2). The subject pronominal form *anh* ‘2SG.Older.M’ in (143a) was therefore appropriately deployed to mark both the age range and the gender of the referent. Similarly, the same pronominal form *anh* was correctly used by Thomas in (143b) to index the same information despite the fact that the grammatical person-reference has changed from 2SG to 1SG.

In contrast, the use of the 3SG *nó* in (144b) below is coded as inappropriate.

- (144) a. Henry<sub>2</sub>: *Ø cũng phải ba mấy bốn chục rồi.*  
 also must three something four ten PERF  
 ‘(He) must be thirty something or forty years old.’  
 [11]
- b. *mà nó nghỉ làm cho thằng Ramsay.* [Inappropriate]  
 but 3SG stop work for guy Ramsay  
 ‘But he-YOUNGER stopped working for Ramsay.’  
 (Tom.Henry.0809, 35:55.6–37:21.9)

In this example, (144a) suggests that the referent is in his late thirties to forties, which is significantly older than Henry, who was 25 at the time of the recording (Chapter 2, Table 2.2). The appropriate form of pronominal form should thus be *chú* (Figure 6.2), but Henry used *nó* in (144b), which indexes a younger referent. The overt form *nó* is thus coded as inappropriate.

<sup>120</sup>Only certain forms are considered because not all types of pronominal form encode gender and age information. Most pronouns (with the exception of 3SG *nó* and 3PL *họ*) or personal names, for example, are generally gender-neutral and age-cue-free.

However, such relevant demographic information of the referent is not always available, and so in these cases, the tokens are excluded from the analysis (N=7).

### 6.3.3 Coding overt pronominal objects

#### 6.3.3.1 Redundant overt pronominal objects

As for the subjects, I first identify overt forms that appear in contexts where a null object pronominal form is typically expected (i.e. a shared structure, a shared verb, or repeated information). These cases are marked as inappropriate. This coding scheme is demonstrated in (145).

- (145) a. Lina<sub>1</sub>: *con lấy cơm ra cho mẹ.* [No pronominal object]  
 2SG.kin take rice out for 1SG.kin  
 ‘You-CHILD take the rice out for me-MOTHER.’
- b. Lina<sub>1</sub>: *con đưa em đây,* [Appropriate]  
 2SG.kin bring 3SG.kin here  
 ‘You-CHILD bring him/her here.’
- c. *mẹ nắm em cho.* [Inappropriate]  
 1SG.kin hold 3SG.kin DM  
 ‘I-MOTHER hold him/her.’

(Lina.Naomi.0623, 14:34.7–14:37.6)

This example features a continuous dialogue between Lina and Naomi. In this example, a natural reading of the full utterance in (145b) and (145c) is that ‘You bring him/her here for me to hold Ø.’ The overt *em* ‘him/her’ (which refers to their dog) after the predicate *nắm* ‘hold’ in (145c) is thus unduly emphatic, especially when the object pronominal form is already present in the immediately preceding utterance. The overt use of object pronominal form here is accordingly marked as inappropriate. On the other hand, (145b) occurs in a context where an overt object is expected (considering that the object in the preceding (145a) is different), and is therefore coded as appropriate.

### 6.3.3.2 Type of pronominal objects

Next, the procedure for marking type of pronominal subjects also applies to type of pronominal objects. Table 6.4 provides an overview of the distribution of different object pronominal types in CanVEC.<sup>121</sup>

	Gen 1		Gen 2	
	N	%	N	%
<b>Overt objects</b>				
<b>Pronouns</b>	68	15%	82	25%
<b>Kin terms</b>	345	74%	233	70%
<b>Personal names</b>	52	11%	19	5%
<b>TOTAL</b>	465	100%	332	100%

Table 6.4: An overview of the distribution of different pronominal types for overt objects in CanVEC. Note that the use of ‘kin term + personal name’ (e.g. *cô Trang* ‘Aunt Trang’) are coded under kin terms in this table.<sup>122</sup>

Similarly to subjects, we see that all three types of pronominal forms are present in the corpus, with kin terms consistently being the most frequent across both generations. To assess whether the choice of each form is appropriate in a given context, consider example (146).

- (146) a. Tanner<sub>1</sub>: ở Đà-Nẵng con có gặp cô Trang không? [Appropriate]  
 LOC Da-Nang 2SG.kin AFF meet 3SG.kin Trang Q  
 ‘Did you-CHILD meet aunty Trang in Đà Nẵng?’

<sup>121</sup>As Theresa Biberauer (p.c.) points out, it is striking that while there is a big difference between the generations in terms of the total number of overt subjects they produce (Gen 1: 2822 vs. Gen 2: 560, Table 6.3), their respective production of overt objects is much closer in number (Gen 1: 465 vs. Gen 2: 332, Table 6.4). The most likely reason is perhaps a combination of genre effects and discourse factors (topic, accessibility of the referents, the distribution of transitive/intransitive verbs in the corpus and so forth). For genre effects specifically, since CanVEC data is highly conversational, first- and second-person subjects can already be understood without being overtly realised. In contrast, objects are less clearly assumed, and hence more often require overt realisation. For example, in conversational Vietnamese, ‘Ø saw him’ is typically understood as ‘I saw him,’ whereas ‘I saw Ø’ is more ambiguous (see also Valian, 2016 for a similar claim cross-linguistically, where subjects tend to index old information and objects new). This fact alone obviously cannot directly explain the numerical differences between generations for overt subjects and overt objects, but it shows that the consequences for dropping objects may be more serious than they are for dropping subjects in conversations. In fact, as we previously saw in Chapter 5, speakers overtly realised objects more frequently than subjects (~80% vs. ~70% respectively), despite far lower overall frequency in discourse. Due to this constraint, speakers may be more consistent in their frequency of object production in order to avoid ambiguity. For subjects, this concern for ambiguity is less relevant, and hence speakers’ production fluctuates more widely owing to discourse factors such as topics and all else.

<sup>122</sup>This is because kin terms play the dominant role of indexing pragmatic information in these cases. For example, to refer to/address Aunt Trang, speakers can say *cô Trang* ‘Aunt Trang’ or *cô* ‘Aunt’ (without Trang), but not *Trang* (without *cô* ‘Aunt’). As such, personal names in these cases only add specificities to the referent, they do not carry pragmatic loads per se. On this basis, I consider ‘kin term + personal name’ as an expression of kin terms only for our purpose.

- b. Nina<sub>2</sub>: *không Ø không gặp Trang.* [Inappropriate]  
 No NEG meet Trang  
 ‘No (I) didn’t meet Trang.’

(Tanner.Nina.0609, 08:19.6–08:35.3)

In this example, Tanner used the correct type of pronominal form *cô* to refer to a female referent who is much older than Nina.<sup>123</sup> His utterance in (146a) is therefore coded as appropriate. In contrast, Nina refers to aunty Trang using personal name only (i.e. without the *cô* in (146b)), which is not expected for reference to seniors. Nina’s utterance is thus marked as inappropriate.

### 6.3.3.3 Form of pronominal objects

For the form of pronominal objects, I again consider whether the correct term has been deployed to denote the right gender and age range of the referent (N=704). Example (147) demonstrates.

- (147) a. Tressie<sub>2</sub>: *con chưa hỏi cô-Heather nhiều.* [✓Gender  
 1SG.kin NEG ask 3SG.kin-NAME much XAge range]  
 ‘I<sub>CHILD</sub> haven’t asked aunt<sub>FEMALE.YOUNGER-THAN-PARENTS</sub> Heather much.’

- b. Harry<sub>1</sub>: *bác-Heather cũng nói giọng Huế.* [Background]  
 3SG.kin-NAME also speak accent Hue  
 ‘Aunt<sub>OLDER-THAN-PARENTS</sub> Heather also has a Hue accent.’

(Harry.Tressie.Josh.0719, 21:57.6–22:00.0)

In this example, Tressie refers to a third person as *cô Heather*, which denotes that Heather is a middle-aged woman who is slightly younger than Harry, Tressie’s parent. Heather, however, is also a speaker in CanVEC, who is known to be five years older than Harry, Tressie’s father. Harry referring to Heather as *bác* in (147b) (which denotes an older middle-aged person) is thus considered more appropriate. Nonetheless, looking at these utterances alone, it remains difficult to judge with any certainty whether Tressie misjudged the age of Heather relative to Harry, or whether she was using the wrong form out of ignorance. In this particularly case, however, further evidence is available later in the transcript (148), Tressie continued to refer to Heather as *cô Heather*, even after Harry’s correction to *bác* in (147b).<sup>124</sup>

- (148) a. Tressie<sub>2</sub>: Tom *cô-Heather* he can’t speak much but...  
 3SG.kin-NAME  
 ‘Tom from aunt<sub>FEMALE.YOUNGER-THAN-PARENTS</sub> Heather(’s family) he can’t speak much but...’

<sup>123</sup>Note that the pronominal forms in Vietnamese are frequently deployed from the younger speaker’s point of view. In this case, for example, Tanner referring to Trang as *cô* does not signify that Trang is older than Tanner (i.e. the father of Nina) but because that is a form that would be appropriate for Nina to use. Theresa Biberauer (p.c.) points out that Tanner’s use of *cô* in this sense is similar to child-directed speech in English, e.g. parents say ‘Aunty Li is going to visit us’ in a context where the child needs to use ‘Aunty Li’ as a form of address that reflects his/her age/status relative to the referent (i.e. Li) rather than it signifying the parents’ own age/status to the referent.

<sup>124</sup>Note that (148) belongs to the code-switching portion, which is not the focus of the investigation in this chapter. However, it is still part of a coherent transcript, which is used to assist the analyses of ambiguous cases as in (147).

- b. Harry<sub>1</sub>: yeah he getting a bit from *bố Phát với-lại mấy* uncle *đó*.  
 father Phat together-with PL DM  
 ‘Yeah he (was) getting a bit (of talking) from father Phát and the uncles.’  
 (Harry.Tressie.Josh.0719, 22:06.3–22:17.2)

This continued use of the inaccurate form suggests that Tressie’s *cô* in (147a) is more likely a product of her uncertainty regarding the use of *cô* and *bác*, rather than a real-world-based knowledge deficit. Tressie’s use of *cô* in (147a) is thus marked as inappropriate in terms of age.

This example also highlights the difficulty that researchers face in analysing appropriate vs. inappropriate use of pronominal forms in Vietnamese. In this work, only cases where some kind of certainty can be established, such as (147), are counted. Ambiguous cases where further evidence is not available to reach a more definite conclusion are excluded (N=2).

### 6.3.4 Coding overt copulas followed by adjectival predicates

For AdjP predicates, null copula is the unmarked choice, while the overt copula *là* marks emphasis. As indicated in Chapter 5 (§5.3.3), an overt copula must co-occur with an intensifier and/or a perfective particle. The example in (149) illustrates:

- (149) Dany<sub>1</sub>: *nó rất là lười*. [Appropriate]  
 3SG INTSF COP lazy  
 ‘He is very lazy.’  
 (Lami.Dany.0825, 01:22.5–01:24.0)

In this example, the copula *là* is realised in an AdjP predicate environment, together with the intensifier *rất* ‘very’, and is therefore coded as appropriate. In contrast, if overt copulas are not accompanied by these structural elements, they are coded as inappropriate, as illustrated in example (150).

- (150) a. Lami<sub>2</sub>: *cái đấy là dễ*, [Inappropriate]  
 CLS DEM COP easy  
 ‘That one is easy.’
- b. *nhưng-mà em-ấy lại không làm được*.  
 but 3SG.kin PRT NEG do ASP.Acquired  
 ‘But he couldn’t do it.’  
 (Lami.Dany.0825, 11:27.3–11:31.2)

Here, *là* occurs before an AdjP predicate without any intensifier or perfective particle. An overt copula is unnecessary in this case and therefore considered inappropriate.<sup>125</sup>

Overall, the coding procedure for overt copulas is more straightforward than for pronominal subjects and objects, as it does not require any language-external resources such as knowledge about age and gender of the referent. The only discourse element here is emphasis, which is also lexically encoded by the combination of an overt copula and a suitable intensifier/perfective particle.

Having described the coding scheme for overt subjects, objects and copulas, I next discuss the results and analysis.

## 6.4 Results

Table 6.5 presents the findings for overt subjects, overt objects, and overt copulas by generation. Note that the total number of each variable provided for each generation is the previously reported total number minus exclusions in each case.

		Redundant			Inappropriate Type			Inappropriate Form		
		N	%	Total	N	%	Total	N	%	Total
<b>Overt subjects</b>	Gen 1	3	0.1%	2,815	1	0.1%	2,815	0	0.0%	1,816
	Gen 2	25	4.5%	560	55	9.8%	560	85	16.7%	510
<b>Overt objects</b>	Gen 1	2	0.4%	463	0	0.0%	463	2	0.5%	390
	Gen 2	3	0.9%	332	3	0.9%	332	3	1.0%	312
<b>Overt copulas</b>	Gen 1	0	0.0%	588	–	–	–	–	–	–
	Gen 2	1	0.3%	296	–	–	–	–	–	–

Table 6.5: Results of Vietnamese overt subjects, objects, and copulas in CanVEC

The first fact that Table 6.5 makes clear is that the number of instances of inappropriate use is generally very low across all variables. For overt objects and overt copulas in particular, no substantial cross-generational difference is observed. The very few cases recorded for both gen-

<sup>125</sup>Note that while the absence of a particle can be compensated for by stressing the copula *là* to achieve emphasis (e.g. *Cái này là đẹp* ‘This one IS pretty’), this construction is rare and often only occurs as part of a serial contrast in Vietnamese (e.g. ‘This one IS pretty, that one IS ugly’). This kind of construction is not found in CanVEC. A more common kind of prosodic compensation for the absence of the copula *là*, however, is where *là* is not realised altogether, but the stress is placed on the adjective itself (e.g. *Cái này đẹp* ‘This one (is) PRETTY’; see also Chapter 5, §5.3.3). This means that an overt copula occurring before an AdjP without any intensifier or appropriate particle is almost always inappropriate in general, and always inappropriate in CanVEC.

erations may well be speakers' errors, and even if not, the numbers are too low to be statistically significant ( $p < 0.01$ ). This converges with the results for null forms in Chapter 5, substantiating the conclusion that objects and copulas remain stable in the Canberra Vietnamese community.

In contrast, there is a sharp discrepancy between the generations for overt subjects. Specifically, while redundant forms of pronominal subjects are almost non-existent for the first-generation, they are substantially more frequent for the second generation, despite a smaller sample size ( $\chi^2 = 103$ ,  $df = 1$ ,  $p < 0.01$ ). This fact is particularly striking in relation to subjects with inappropriate types (0.1% vs. 9.8%) and inappropriate forms (0.0% vs. 16.7%). The results again mirror the trend found in Chapter 5, where there is a cross-generational difference for subjects, but not objects and copulas.

Having established where cross-generational effects are observed for overt forms, we next take a closer look at each of the cases where the effects are detected.

### 6.4.1 Redundant overt subjects

Looking at all the cases of inappropriate overt pronominal subjects in more detail, we observe that all the inappropriate pronominal forms produced by the first generation come from Reece, the oldest speaker in the corpus (68 years old at the time of the recording). These instances are reported below. The preceding turn from the second-generation speaker in (151) is also given for context.

(151) a. Taylor<sub>2</sub>: *họ mua sẵn cái gì?*  
 3PL buy ready-made CLS what  
 'What ready-made stuff did they buy?'

b. Reece<sub>1</sub>: *họ mua sẵn những cái giấy đó.*  
 3PL buy ready-made PL CLS paper DM  
 'They bought some ready-made paper.'

(Reece.Taylor.0906, 06:18.1–06:20.8)

(152) a. Reece<sub>1</sub>: *nó nói tiếng Anh đó,*  
 3SG speak language England DM  
 'She speaks English,'

b. *nó nói dữ lắm.*  
 3SG speak much INTSF  
 'She speaks a lot.'

(Reece.Taylor.0906, 09:36.4–09:44.6)

(153) a. Reece<sub>1</sub>: *cô Mi chưa biết tiếng Anh,*  
 3SG.kin Mi NEG know language England  
 'Aunt Mi did not know English,'

- b. *cô Mi chưa biết đọc.*  
 3SG.kin Mi NEG know read  
 ‘Aunt Mi did not know how to read.’

(Reece.Taylor.0906, 41:00.0–41:03.5)

In all these cases, the subject pronominal forms are realised in constructions where the VPs are repeated from the preceding clause, rendering the overt forms redundant. Given that all of these isolated instances come from Reece, this pattern is not representative of the first-generation speakers. As for why Reece in particular produces these instances, we may speculate that it is because of Reece’s more advanced age,<sup>126</sup> or because of his English-dominated conversation (which is also the most English-dense transcript in the corpus, ~70%, N=674/962 clauses). Limited data, however, precludes us from drawing a conclusion.

In contrast, 12 out of 17 second-generation speakers produce at least one redundant overt pronominal subject, which suggests this is much more of a characteristic—if an infrequently attested one—for this generation than for the first generation. Examples (154)–(156) below exemplify some instances. The preceding turns from the first-generation speakers are also provided for context.

- (154) a. Tanner<sub>1</sub>: *con nhớ Hội-An không?*  
 2SG.kin remember Hoi-An NEG  
 ‘Do you-CHILD remember Hoi-An?’

- b. Nina<sub>2</sub>: *con có nhớ Ø,*  
 1SG.kin AFF remember  
 ‘I-CHILD remember (it),’

- c. *mà con cũng không có nhớ Ø.*  
 but 1SG also NEG AFF remember  
 ‘But I-CHILD also don’t remember (it).’

(Tanner.Nina.0609, 04:24.3–04:43.2)

- (155) a. Mina<sub>1</sub>: *ủa bố lấy Ø đi đâu?*  
 DM 3SG.kin take go where  
 ‘Oh where did he-FATHER take (it) to?’

- b. Pete<sub>2</sub>: *bố tặng Ø cho bạn.*  
 3SG.kin give for friend  
 ‘He-FATHER gave (it) to his friend.’

(Mina.Pete.0906, 11:02.9–11:07.9)

<sup>126</sup>Although we do not yet have any evidence for a correlation between the age factor and the (over)use of overt pronominal forms, see Kaltsa, Tsimpli & Rothman (2015) for some preliminary connection between age effects and pronoun resolution in Greek. Specifically, they ran a self-paced sentence-picture matching experiment on 91 speakers of Greek (both monolinguals and bilinguals) and found that the older participants (range: 55–65) seem to favour matching an overt pronominal form to a subject antecedent more than the younger participants (range: 19–34). This tendency is somewhat reflected in Reece’s production in (151)–(153), where all the instances of his overuse of overt pronominal forms have subject antecedents.

- (156) a. Penny<sub>2</sub>: *con lấy tay má xong,*  
 1SG.kin take hand 2SG.kin PERF  
 ‘I-CHILD took your-MOTHER hand first.’
- b. *con từ-từ đi bộ qua đường.*  
 1SG.kin slowly go foot across road  
 ‘(then) I-CHILD slowly walked across the road.’

(Penny.Marie.Rory.0912, 11:39.1–11:42.8)

In all these cases, although the VPs are not always repeated as they were for Reece (examples (151)–(153)), the structures between turns are highly similar. Since the subject referent is also continuous and immediately accessible from discourse, the repeated overt pronominal forms come across as unnecessarily emphatic. Overall, the observation therefore supports the previous findings on the over-extension of heritage overt subject pronominal forms to contexts where null forms are typically expected (§6.2.1).

## 6.4.2 Pronominal type and pronominal form

As for pronominal type and pronominal form, there is some significant overlap between their findings. As Table 6.5 shows, there are 55 cases of inappropriate pronominal form use for the second generation. On closer inspection, I find that 100% of these cases involve the choice of a 3SG gender-neutral pronoun *nó* over a kin term indexing a specific gender. Example (157) illustrates this.

- (157) a. Luna<sub>1</sub>: *anh Roland hồi xưa khó tính.*  
 3SG.kin Roland time old difficult personality  
 ‘Brother-OLDER-THAN-TRESSIE Roland used to be difficult.’
- b. Tressie<sub>2</sub>: *bây-giờ nó dễ rồi.*  
 now 3SG easy PERF  
 ‘He-YOUNGER is easy now.’

(Luna.Tressie.0901, 09:24.8–09:29.2)

- (158) a. Marie<sub>1</sub>: *chị đó là quản-lí chuỗi nhà-hàng đó.*  
 3SG.kin DEM COP manager chain restaurant DM  
 ‘She-OLDER-THAN-PENNY is the manager of a restaurant chain.’
- b. Penny<sub>2</sub>: *nó giàu vậy hả?*  
 3SG rich DEM Q  
 ‘Is she-YOUNGER that rich?’

(Penny.Marie.Rory.0912, 05:02.0–05:05.1)

In (157), the referent is Roland, who Luna (a first-generation speaker) refers to as *anh*. This suggests that the speaker is a known acquaintance/member of the family who is older than

Tressie.<sup>127</sup> In response, however, Tressie refers to Roland using a gender-neutral pronoun *nó*. Similarly, Penny also used the pronoun *nó* in (158b) in place of *chị* (158a). This type of pronoun choice is pragmatically inappropriate in both cases, especially when referring to a senior (Chapter 5, §5.3.1).

Strikingly, the misuse of *nó* also accounts for 65% (N=55/85) of all the instances of misused pronominal subjects. Specifically, *nó* (which indexes a younger referent) is frequently deployed where a kin term indexing an older referent is expected. Table 6.6 highlights the prominence of this tendency. Similarly, *bác*, another gender-generic form that indexes a referent older than the speaker's parents, is also frequently used where a kin term indexing a younger referent is expected. This pattern of over-generalisation resembles what we saw for the general classifier *cái* in Chapter 4. In all cases, it is the generic variant that extends its use into domains where a more specific, nuanced form is typically expected.

Form used	Form intended	N	%
<i>nó</i> (3SG)	<i>anh</i> (3SG)	30	35%
Male/Female	Male		
Younger than speaker	Older than speaker		
<i>nó</i> (3SG)	<i>chị</i> (3SG)	25	30%
Male/Female	Female		
Younger than speaker	Older than speaker		
<i>bác</i> (3SG)	<i>chú</i> (3SG)	15	18%
Male/Female	Male		
Older than parents	Younger than parents		
<i>bác</i> (3SG)	<i>cô</i> (3SG)	9	11%
Male/Female	Female		
Older than parents	Younger than parents		
<i>cô</i> (3SG)	<i>bà</i> (3SG)	3	4%
Female	Female		
Younger than parents	Grandparents' age		
<i>chị</i> (3SG)	<i>cô</i> (3SG)	2	2%
Female	Female		
Older than speaker	Younger than parents		
<b>TOTAL</b>		<b>85</b>	<b>100%</b>

Table 6.6: The distribution of inappropriate overt pronominal forms. With the exception of *nó* (3SG pronoun), all the terms here are kin terms.

<sup>127</sup>Recall that the pronominal form used is frequently deployed from the younger speaker's point of view.



- b. Tressie<sub>2</sub>: *mai-mốt me phải tập con rồi.* [1SG object – Appropriate]  
 later 2SG.kin must train 1SG.kin PRT  
 ‘You-MOTHER have to teach/train me-CHILD later.’  
 (Luna.Tressie.0901, 04:22.4–04:25.4)
- (161) a. Tyler<sub>2</sub>: *anh kể em bao-giờ chưa nhi?* [2SG object – Appropriate]  
 1SG.kin tell 2SG.kin ever IMPERF Q  
 ‘Have I-MALE.OLDER ever told you-MALE/FEMALE.YOUNGER?’
- b. Ellie<sub>1</sub>: *em nhớ.*  
 1SG.kin remember  
 ‘I-MALE/FEMALE.YOUNGER remember (you did).’  
 (Billy.Tyler.Ellie.0807, 28:50.6–28:52.5)<sup>128</sup>

Considering that self- (1SG) and interlocutor- (2SG) reference are familiar and frequently present in discourse, this finding is intelligible. Given the close relationships between the speaker and the interlocutor in this study (Chapter 3, §3.2.1), 1SG and 2SG references are likely already firmly established in the speakers’ lexicon. In other words, there is more discourse and pragmatic information available to allow the speakers to select an appropriate pronominal type/form for 1SG and 2SG reference.

This skewed distribution of grammatical person, however, is not observed for the second generation’s use of overt subjects. In fact, all grammatical persons are distributed more equally in this domain, with 3SG accounting for over 30% of the second generation’s total production (N=163/510). This is significantly higher than the corresponding figure for objects, both numerically and proportionally. Given that all the problematic cases fall into 3SG, the identified discrepancy between the second generation’s overt subject misuse and overt object misuse seems most likely an artefact of data distribution.

For overt copulas, Table 6.5 shows that this is where the least variation is observed, in terms of both the raw count of inappropriate cases and the cross-generational differences. In fact, first-generation speakers produce no cases of inappropriate overt copulas, while second-generation speakers produce only one, reported in (162).

- (162) a. Lami<sub>2</sub>: *cái đấy là dễ,*  
 CLS DEM COP easy  
 ‘That one is easy.’

<sup>128</sup>Recall from Chapter 2 that generation membership is not necessarily age-correlated in the context of this study. In (161), for instance, Tyler is classified as a Gen 2 speaker because he is an Australian-raised early bilingual, while Ellie is classified as Gen 1 speaker because she is a late bilingual arriving as an adult after 18. However, Tyler (Gen 2) is still older than Ellie (Gen 1) (Chapter 2, Table 2.2) and as such the pronominal forms they use here are appropriate.

- b. *nhưng-mà em-ấy lại không làm được.*  
 but 3SG.kin PRT NEG do ASP.Acquired  
 ‘But he couldn’t do it.’

(Lami.Dany.0825, 11:27.3–11:31.2)

In this example, the copula *là* is realised with an AdjP predicate, but without any intensifier or a perfective particle. The realisation of overt copula as it is in this example is inappropriate.

Setting this isolated case aside, the fact that overt copulas are so well-preserved, particularly among the second generation, is especially noteworthy. Although this result contrasts what we saw for Spanish *estar* (§6.2.1), it is consistent with what was expected for internal interface phenomena (§6.2.2), namely that they should be rather stable properties. This observation supports a distinction between internal-interface components and external-interface components, with the former appearing to be more successfully retained in bilingual situations.

## 6.5 Chapter summary

In this chapter, I explored the usage of overt forms of subjects, objects, and copulas in the heritage Vietnamese of the speakers in the CanVEC corpus. In terms of cross-generational variation, results converge with what was found for null forms: the generational effects are only observed for subjects, while objects and copulas remain stable. Findings also concur with findings in the previous literature in that:

- (i) the distribution of overt pronominal subjects is often extended for heritage language speakers of all kinds (i.e. both generations in this study); and
- (ii) early-bilingual heritage language speakers (Gen 2) deviate more in domains concerning subtle nuances of semantics-pragmatics shades (i.e. the age index of pronominal forms in this case) than those with categorical elements (i.e. the gender index of pronominal forms in this case).

In the final chapter of the thesis, [Chapter 7](#), I bring all the findings from [Chapters 4–6](#) together and discuss implications for future work.

---

## CONCLUDING REMARKS

---

I began this dissertation with two main objectives:

- (i) to document the vernacular of the Canberra Vietnamese community; and
- (ii) to consider key aspects of the vernacular of this community in relation to cross-generational variation from prominent theoretical perspectives on data of the relevant kind.

The first objective was the focus of Part I of this work (Chapters 2–3), where I described the Canberra Vietnamese-English corpus (CanVEC), an original dataset that was newly compiled for this study. The corpus features over 10 hours of natural speech from 45 speakers across two generations in the community. The significance of this resource has been discussed at several points in this dissertation; that is, it represents the first annotated and freely available corpus of the speech of the Canberra Vietnamese community.<sup>129</sup> The second objective, which is to put the corpus to use and characterise cross-generational variation in the community, was the focus of Part II (Chapters 4–6). My approach in this part involved an integrated perspective from three different theoretical frameworks: the Matrix Language Turnover Hypothesis (based on the Matrix Language Framework) (Myers-Scotton, 1993 et seq.), the variationist framework (Labov, 1972 et seq.), and the generative interface vulnerability approach (Sorace & Filiaci, 2006 et seq.), each of which was deployed at a different stage to unpack different aspects of the data in question. By integrating different theoretical standpoints in studying this newly collected dataset, this dissertation contributes a multi-faceted treatment of a heritage language in a community that has not been previously examined.

My aims in this concluding chapter are three-fold:

- (i) to summarise the key findings in relation to heritage Vietnamese in the Canberra Vietnamese bilingual community (§7.1);

---

<sup>129</sup><https://github.com/Bak3rLi/CanVEC>

- (ii) to discuss the broader implications of these findings for heritage languages in general (§7.2); and
- (iii) to highlight specific questions that my dissertation has directly raised, as well as further possibilities for future research (§7.3).

## 7.1 Heritage Vietnamese in the Canberra bilingual community

The investigation of cross-generational language variation and shift began in [Chapter 4](#), where I first probed the code-switching production of the corpus, using the influential MLF and the associated ML Turnover Hypothesis (Myers-Scotton, 1998). The ML Turnover Hypothesis predicts that when the original Matrix Language, i.e. the language that provides the morphosyntactic frame for a bilingual CP, becomes the Embedded Language in the community, structural borrowing in the direction of the new Matrix Language will follow. The CanVEC data showed, however, that this prediction is insufficiently nuanced. In fact, even when the evidence for the ML Turnover was quantitatively present in the direction of English, abstract structural influence was observed in the opposite direction in the community. Specifically, a large proportion of English sentences from both first- and second-generation speakers were found to contain null arguments or null functional elements, a feature that is permissible in Vietnamese but not typically in English. Furthermore, while novel elements such as articles expressing definiteness were nowhere to be found in monolingual Vietnamese clauses, a handful of otherwise-English clauses (mostly from second-generation speakers) were found to contain the Vietnamese generic classifier *cái*. These patterns together suggest some abstract influence from Vietnamese onto speakers' English, rather than the other way round. Considering this in the context that the MLF could only account for less than half of the speakers' production (42% and 44% for Gen 1 and Gen 2 speakers respectively), I reached the conclusion that the definitions of the MLF component parts are insufficiently clear, and even if one tries to sensibly flesh out these components, the ML Turnover Hypothesis gives predictions that do not seem to reflect what we see in CanVEC.

Faced with this difficulty, I turned to the variationist approach (Labov, 1972) in [Chapter 5](#) to further probe an area that the MLF particularly struggled to account for: null elements. In this chapter, I moved away from the CanVEC code-switching subset to probe the heritage Vietnamese monolingual subset directly. Specifically, I compared cross-generational patterns relating to three cases where the null and overt alternation exists in Vietnamese: subjects, objects, and copulas. Variationist results offer some illuminating insights into cross-generational variation in the community: while the patterns of null objects and null copulas remained stable, changes were detected for null subjects. In particular, first-generation speakers were more likely to drop

first-person subjects, while second-generation speakers were more likely to drop second-person subjects. Given the complex Vietnamese honorific system, which requires second-generation speakers to always overtly realise forms referring to their interlocutors, this specific direction of effects runs counter expectation. Considering the community background and consistent patterns across the corpus, I explained this finding in terms of cultural integration into the Australian society. Specifically, I considered this linguistic pattern a possible form of community bricolage (Eckert, 2004) that rejects the entrenched social hierarchy in the heritage language, thereby establishing a more equal relationship between the generations. In this sense, my analysis here supports the so-called **third-wave** sociolinguistics emphasis on micro-level social interactions and individual identity as primary forces driving change. In the context of the Canberra Vietnamese community, the consistency among speakers of the same generation further indicated that this behaviour has gained traction and become an established pattern within the community.

Another key observation that emerged from [Chapter 5](#) is that across all the three variables of interest (subjects, objects, and copulas), speakers overwhelmingly prefer overt forms over null forms (~70% of total production in their Vietnamese output). Given that it has frequently been suggested that the overt counterparts of null forms exhibit distinctive behaviour in bilingual contexts, these overt forms became the focus of [Chapter 6](#). In particular, I appealed to the interface-oriented approach (Sorace & Filiaci, 2006; Sorace & Serratrice, 2009a; Sorace et al., 2009b; Sorace, 2011; Tsimpli, 2014; Sorace, 2016) to establish whether the different interface factors regulating the occurrence of overt subjects, objects and copulas in colloquial Vietnamese were preserved in the Canberra community, or whether this community also exhibits interface vulnerability effects similar to those that have been uncovered in other bilingual communities. Results showed that for overt objects and overt copulas, no substantial cross-generational difference was observed, while for overt subjects, there was a clear discrepancy between the generations. Specifically, second-generation speakers used inappropriate pronominal types (e.g. pronouns in place of kin terms) and inappropriate pronominal forms (e.g. those that index inaccurate age ranges) significantly more frequently than first-generation speakers. These results converge with the results for null forms in [Chapter 5](#), substantiating the conclusion that cross-generational difference is only observed for Vietnamese subjects, but not objects and copulas in the Canberra Vietnamese bilingual community.

## 7.2 Heritage languages in a broader context

Tying these results together, there are two distinct points that can be made, not just about Vietnamese in Canberra, but also about heritage languages in general. First of all, it is clear that the cross-generational effects in Vietnamese heritage language are property-specific: changes are observed in some, but not in all linguistic phenomena. This tendency is consistent with what is expected in heritage languages across the board more generally (Montrul, 2015; Polinsky, 2018; Aalberse et al., 2019; Polinsky & Scontras, 2020). In particular, findings highlight the roles of timing of acquisition and support a distinction between the internal and external interfaces among second-generation speakers, whereby internal-interface phenomena are more robustly preserved under contact than their external-interface counterparts. Furthermore, my analyses also highlight the universal tendency of second-generation speakers performing better with categorical one-to-one mapping (e.g. gender index (M/F/N)) over variable one-to-many mappings (e.g. relative age index) in pronominal forms. Given that the question of what remains stable and what changes in heritage languages has already become a central focus in the field of heritage languages in recent years (e.g. Aalberse et al., 2019; Polinsky & Scontras, 2020), results here have only highlighted the need to further probe this area.

It is worth noting, however, that these seemingly universal patterns of heritage languages in general do not mean that heritage language is a homogeneous ‘type’ of language, spoken by a homogeneous group of speakers. As Chapters 4–6 collectively demonstrate, any single analysis that relies on only the grammatical aspects without taking into account the nuanced intricacies of acquisition, the community, and the specificity of the varieties involved, runs the very real risk of failing to highlight systematic patterns that can afford us a deeper understanding of what is at work. In this study, for example, while the pattern of use is highly consistent among speakers within their generation, I illustrated near the end of Chapter 5 that there is in fact a fair amount of variance between speakers in terms of their rates for pronominal subject drop (45-95% for first generation’s most preferred subject drop (i.e. first-person), and 30-82% for second generation’s most preferred subject drop (i.e. for second-person)). Although this is possibly partly determined by topic and discourse factors, the huge variance observed nevertheless highlights the fact that individuals within a community may behave in quite different ways, depending upon which aspect of their production is being examined. This, together with the fact that some speakers do not just differ in rates but also in their patterns of subject drop, further underscores the complexity and diversity of linguistic behaviour within a heritage language community.

### 7.3 Where to from here?

Having discussed the findings and their implications, I would now like to highlight some priorities for future research. The first concerns the nature of argument drop in Vietnamese and in radical pro-drop languages in general. Specifically, while subject and object drop is said to be rather liberal in radical pro-drop languages, in reality, the drop of these elements might not be so ‘radical’ after all. For heritage Vietnamese, there is a clear tendency for speakers to prefer overt over null forms, with overt realisation accounting for at least 70% of speakers’ Vietnamese production. Although part of this high proportion might be attributed to ‘extended use’ of overt forms in contact scenarios, this dissertation has also highlighted several pragmatic constraints that prevent speakers from dropping subjects even in monolingual varieties. As I made a finer distinction between discourse and pragmatics in [Chapter 6](#), it is worth pointing out that, to date, most of the existing work on pro-drop languages has only focused on the former, i.e. the discourse conditions that regulate null subject realisation (e.g. coreferentiality, ambiguity, distance from the previous mention, etc.; see e.g. Owens et al., 2013; Travis & Lindstrom, 2016; Torres Cacoullos & Travis, 2018; Frascarelli, 2018, i.a.). This trajectory represents a gap in our understanding of Vietnamese-type pro-drop languages, where the pragmatic factors such as politeness, interlocutor’s age and their perceived social status are considered to be just as important ([Chapters 5–6](#)). Given that the division of labour between null and overt arguments in radical pro-drop languages in general is still only partially understood (Jia & Bayley, 2002; Ngo, 2019), probing these specific pragmatic elements may be crucial in shedding some light on this sparse area of research. Advances in our understanding of this domain will also in turn allow us to further identify the universal and language-specific areas that are most likely to be vulnerable under contact.

In the broader context of language variation and change, this dissertation has also highlighted the need for more acquisition work, particularly on minority languages. As [Chapter 4](#) demonstrated, the acquisition angle offered a valuable insight into the over-generalisation of the generic classifier *cái* in relation to the connection between timing of acquisition and the stability of certain syntactic knowledge over time, even in scenarios of sustained contact. This aspect of acquisition also consistently emerged as relevant in [Chapter 5](#) and [Chapter 6](#), where results converged to support the prediction that the earlier a property is acquired, the more likely it is to remain present in a speaker’s repertoire, independent of change in input. Unfortunately, however, since work on acquisition in Vietnamese and in minority languages in general is still extremely limited, we do not yet know much about this area with respect to various properties. In my view, this should be a priority for future research.

Beyond these broad matters that have general implications for the field, I would also like to foreground some specific points and propose possible extensions to my work that have implications for Vietnamese in particular. The first concerns the Labovian idea of a ‘speech community,’ which traditionally attaches weight to an established **speech norm** within a group. As I highlighted in [Chapter 2](#), however, given that sociolinguistic work on heritage Vietnamese is sparse and evidence for a defined Canberra speech norm is almost non-existent, applying this definition to an atypical community like the Vietnamese in Canberra is not so straightforward. Findings in the second part of this dissertation, however, made some progress towards this by uncovering established speech patterns within and across generations in the community, not only in their code-switching discourse ([Chapter 4](#)), but also in their monolingual English ([Chapter 5](#)) and monolingual Vietnamese (Chapters 5–6). This ultimately strengthens support for a cohesive Canberra Vietnamese speech community, and provides a baseline against which further comparative work can be conducted. With this groundwork now in place, I suggest that future research focuses on other heritage Vietnamese communities in Australia and elsewhere, while harnessing the tools that I introduced in [Chapter 3](#) to create comparable corpora. Once these resources are in place, comparison between different heritage Vietnamese varieties, and between heritage Vietnamese and other low-resource, minority heritage language varieties will become possible.

Similarly, documenting spoken Vietnamese data from the homeland should also be actively investigated. Although written Vietnamese is generally accessible, spoken Vietnamese is still extremely difficult to obtain. I pointed out at several places in this dissertation (Chapters 3–5) that written Vietnamese and spoken Vietnamese are quite different and that this fact needs to be properly recognised. Unfortunately, work on Vietnamese thus far has mainly relied on existing materials in written Vietnamese only, which is not always an appropriate benchmark for various phenomena (e.g. pronoun use or discourse markers of politeness). In this study, while I was fortunate to have access to some short recordings (Vietlex data, [Chapter 5](#)) and a small corpus of spoken Southern Vietnamese (Brunelle’s data, Chapters 5–6) as points of reference, more collective efforts are sorely needed to increase the availability and accessibility of colloquial Vietnamese as spoken in the homeland. Only when such a resource is available, can we seriously investigate whether modern spoken Vietnamese varieties also exhibit any of the properties found in heritage Vietnamese as reported in this work.

As for code-switching research in general, specific attention should also be paid to Vietnamese and other under-studied language pairs. The work I presented in [Chapter 4](#) particularly shows how challenging it can be to investigate a language pair that has rather limited overt morphology and a similar clausal word order, i.e. structural elements that are believed to ‘frame’ the gram-

mar of a mixed clause. Based on data from CanVEC, I concluded that code-switching cannot and should not be analysed as surface combinations of two prescriptively sanctioned monolingual standards. The question, however, remains open as to how else we can best approach this challenging dataset. As I stressed the general need to factor in timing of acquisition, the nature of the community, and the linguistic specificity of the varieties under study, I rely on future work to incorporate all these relevant elements.

For Vietnamese-English code-switching more specifically, it should also be noted that although some research was conducted many years ago (Tuc, 2003), data is no longer available and virtually no active research on this topic has been recorded ever since (see, however, Nguyen, 2018). What this dissertation now provides, then, is a readily usable resource and some preliminary insights to revive interest in this particular language combination. As I made a case in [Chapter 4](#), the existing data on Vietnamese-English code-switching, however sparse, has already challenged several ‘universal’ assumptions such as the asymmetrical structure in bilingual speech and the prohibited switch between a pronoun and a verb in given mixed clauses (Nguyen, 2018). This thus led me to believe that more focused research-efforts in this domain might lead to even more illuminating findings in relation to the nature of code-switching in general.

On the computational front, some minor observations made in this study are also potentially helpful for future research. First of all, [Chapter 3](#) made clear the difficulties that existing Vietnamese POS taggers had on a spoken dataset, especially in relation to Vietnamese pronouns. This is related to the point I previously made about the difference between the written and spoken register of this language, thereby signalling a need for better training data in the spoken domain. As speech technology continues to develop, this need will only become ever more relevant. I also observed near the end of the same chapter that machine translation might perform better with Vietnamese pronouns in a mixed clause than in a monolingual clause, possibly because the better-resource participating language in code-switching (i.e. English) somehow contributed to enhancing the accuracy in the lower-resource language (i.e. Vietnamese). This observation potentially opens up the opportunity to leverage better-resource languages to enhance the translation performance of lower-resource languages in code-switching discourse. Some semi-related work such as translation via intermediate languages is already underway (Sennrich & Zhang, 2019), and future experiments can explore this avenue further.

Finally, although I primarily focused on the broad topic of cross-generational effects in this study, I also want to highlight several opportunities for a more fine-grained analysis. Since CanVEC data is also already fully transcribed and freely available, it is furthermore well-placed to exploit this possibility. For example, I suggested near the end of [Chapter 4](#) that the over-generalisation the Vietnamese generic classifier *cái* has the potential to ultimately feed into a

long-term change in contact scenarios. A closer investigation of this pattern will hence not only provide a solid data point for future diachronic research on Vietnamese heritage language, but may also shed light on the little-understood role of early-bilingual heritage language speakers as potential drivers of change. Additionally, given that CanVEC data consists of spontaneous speech recorded by community members themselves, detailed conversation analyses may effectively reveal further nuances of speakers' local accommodation, which, under the right set of circumstances, may pave the way for long-term accommodation and trigger sustainable language change (Trudgill, 1986; Hinskens & Auer, 1997; Sachdev & Giles, 2004). CanVEC data can also be used to study other dimensions of the Canberra Vietnamese vernacular that have never received any attention in the literature, such as sound change and L1 attrition, as well as comparative work between Vietnamese and other diasporas around the world. In Australia specifically, the possibilities for meaningful linguistic comparison with relevant communities are numerous, as usable data from other heritage language communities is already available (Clyne, 2003) and has become increasingly so in recent years.<sup>130</sup>

As I reach the end of this dissertation, I would like to reiterate the same sentiment with which I began it: the importance of studying under-described heritage languages. Heritage language is still an emerging field, and, as Polinsky & Scontras (2020, p.13) remarked in their recent keynote paper, 'we have barely scraped the surface' of its rich empirical landscape. There are many outcomes and many contact scenarios that the field has not had the opportunities nor the resources to fully explore. The lack of data from a diverse source of communities and language varieties has only contributed to this problem. By focusing on a heritage language in a community that has not been previously examined, this dissertation is thus a contribution not only to the vast body of existing literature on language variation and change, but also to the relatively new field of research on heritage language. My hope is that my creation of CanVEC, and my attempt to probe a lesser-described heritage language such as Vietnamese in an atypical community such as Canberra, will serve as an effective launch pad for sustainable progress in this area.

---

<sup>130</sup>See <http://www.dynamicsoflanguage.edu.au/sydney-speaks/>

---

# INVITATION LETTERS TO PARTICIPATE

---

This appendix presents the bilingual invitation letters that I sent to potential participants at the beginning of the data collection process (Chapter 2, §2.4.1).

## A.1 Vietnamese version

### Giao tiếp song ngữ trong cộng đồng người Việt tại Úc

Tôi viết thư này để mời bạn tham gia vào dự án nghiên cứu khoa học về hội thoại song ngữ trong cộng đồng người Việt ở Canberra. Tôi là nghiên cứu sinh tiến sĩ tại Đại học Cambridge.

Tôi cần bạn ghi âm một hoặc nhiều đoạn hội thoại tự nhiên với người thân hoặc một người bạn song ngữ của mình. Bạn có thể tự chọn người cùng tham gia hội thoại. Tổng thời gian 30 phút (hoặc lâu hơn), và sau đó bạn sẽ được yêu cầu điền vào một bảng câu hỏi ngắn. Kết quả của công trình nghiên cứu này sẽ nâng cao hiểu biết về hành vi ngôn ngữ và sinh hoạt cộng đồng của người nhập cư tại Úc. Bạn sẽ được trả 40 đô la cho đóng góp của mình. Quy trình bảo mật của nghiên cứu này đã được phê duyệt bởi Ủy ban Đạo đức của đại học Cambridge. Các thông tin cần thiết sẽ được cung cấp cho bạn trước khi bắt đầu.

Nếu bạn có hứng thú, xin vui lòng liên hệ nhxxx@cam.ac.uk, hoặc 0432 xxx xxx.

Li Nguyễn

## A.2 English version

### **Vietnamese bilingual communication in Canberra**

I am writing to invite you to participate in a research project on how bilingual Vietnamese communicate with each other in Canberra, Australia. I am a Vietnamese PhD student at the University of Cambridge, funded by the Cambridge International and European Trust.

What I would like to do is to obtain one or many self-recordings of you having an informal conversation with a bilingual member of your family or a friend. You are welcome to choose the bilingual person you would like to be recorded with. The total recording time should be at least 30 minutes (though the longer it is, the better) and you will be asked to fill in a short questionnaire afterwards. The findings hope to provide useful insight into Vietnamese migrants' linguistic behaviour and community practice.

You will be paid 40 dollars for compensation of your time and contribution. The ethical aspects of this research have been approved by the Ethics Committee of the Cambridge Faculty of Medieval and Modern Languages. Further information will be given to you in form of an information sheet before you commence our study.

If you are interested, please contact me by email at nhxxx@cam.ac.uk, or by phone on 0432 xxx xxx.

I look forward to talking to you soon,

Li Nguyen

---

## CANVEC SCORES ON LANGUAGE ATTITUDE

---

This appendix presents the statistics of CanVEC speakers' responses to each pair of adjectives in the questionnaire ([Appendix E](#)), which are designed to measure their language attitude ([Chapter 2, §2.4.4](#)).

	Useful				Friendly			
	Vietnamese		English		Vietnamese		English	
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
<b>Gen 1</b>	3.3	0.3	3.4	0.6	4.2	0.2	2.8	0.7
<b>Gen 2</b>	3.0	1.0	3.2	0.6	3.2	0.5	3.3	1.0

	Inspiring				Beautiful			
	Vietnamese		English		Vietnamese		English	
	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>	<i>Mean</i>	<i>SD</i>
<b>Gen 1</b>	2.5	0.7	3.0	1.0	1.7	0.6	1.7	0.5
<b>Gen 2</b>	1.2	0.1	1.5	0.5	1.0	0.3	2.4	0.9

Table B.1: CanVEC speakers' responses to each pair of adjectives describing Vietnamese and English on a scale from 1 to 5, with 1 being the least and 5 the most positive.



# INFORMATION AND CONSENT FORM

This appendix presents the bilingual information and consent forms, given to participants prior to recordings taking place ([Chapter 3, §3.2.1](#)).

## C.1 Vietnamese version

**Nghiên cứu sinh:** Li Nguyễn, Sinh viên hệ Tiến Sĩ, chuyên ngành Ngôn Ngữ học, Trường Đại Học Cambridge, Vương Quốc Anh.

**Đề tài:** Giao tiếp song ngữ trong cộng đồng người Việt tại Úc.

**Mô tả và phương pháp:** Giao tiếp song ngữ trong cộng đồng người Việt tại Úc. Đề án nhằm nâng cao hiểu biết về việc giao tiếp trong môi trường đa văn hoá.

**Tình nguyện viên:** Đề án tập trung nghiên cứu đối tượng là người Việt Nam thường xuyên sử dụng cả tiếng Anh và tiếng Việt, đã sinh sống và định cư ở Canberra ít nhất 10 năm (thế hệ di dân thứ 1), hoặc người Việt Nam sinh ra ở Canberra hoặc đến Canberra trước 5 tuổi. Người Việt mới dọn đến hoặc chưa ở đây đến 10 năm, hoặc theo gia đình sang Canberra sau 5 tuổi không nằm trong phạm vi nghiên cứu của đề án này. Số liệu được thu thập thông qua băng ghi âm hội thoại và bảng câu hỏi. Đề án chỉ tìm kiếm tình nguyện viên trên 18 tuổi, nhưng nếu trẻ em dưới 18 tuổi cũng tham gia vào đoạn hội thoại, phải có sự cho phép của cả đứa trẻ và bố mẹ/ người bảo hộ để sử dụng số liệu này.

**Sử dụng số liệu:** Kết quả nghiên cứu của công trình này sẽ được báo cáo trong một luận văn tiến sĩ, và có thể sẽ được trình bày tại một hội thảo hoặc dưới dạng một bài báo khoa học. Cô/ bác/ anh/ chị có thể yêu cầu một bản copy khi kết quả hoàn thành.

**Cô/ bác/ anh/ chị sẽ làm những gì trong đề án này?** Cô/ bác/ anh/ chị sẽ được yêu cầu ghi âm một đoạn hội thoại tự nhiên (khoảng 30 phút) và vào một bảng câu hỏi ngắn sau đó. Đoạn hội thoại sẽ được ghi âm và tường thuật lại. Nghiên cứu sinh sẽ không có mặt, cô/ bác/ anh/ chị có thể tự quyết định thời gian và nội dung của cuộc hội thoại mà cô/ bác/ anh/ chị muốn chia sẻ (bảng ghi âm hoàn thành trong thời gian 2–3 tuần). Bảng câu hỏi sau đó sẽ được gửi qua email, hoặc bưu điện, tùy vào lựa chọn của cô/ bác/ anh/ chị. Thời gian tổng cộng cô/ bác/ anh/ chị đầu tư vào dự án này là 50–60 phút.

**Nguy cơ và tính bảo mật thông tin:** Không có nguy hại lớn nào cho cô/ bác/ anh/ chị trong dự án nghiên cứu này. Mục đích của đề án này là tìm hiểu các hành vi ngôn ngữ trong cộng đồng người Việt ở Canberra, chúng tôi không có ý định thay đổi hành vi của cô/ bác/ anh/ chị dưới bất cứ hình thức nào. Tham gia vào dự án này hoàn toàn mang tính tự nguyện, cô/ bác/ anh/ chị có thể rút lui hoặc từ chối trả lời bất kì câu hỏi nào mà không cần giải thích lí do. Nếu cô/ bác/ anh/ chị quyết định rút lui, các số liệu đã thu thập từ cô/ bác/ anh/ chị sẽ bị tiêu hủy, trừ khi cô/ bác/ anh/ chị muốn và cho phép chúng tôi sử dụng.

Tất cả các thông tin cá nhân sẽ được loại bỏ khỏi các tài liệu thu thập được (bao gồm băng thu âm và bản câu hỏi) trước khi số liệu được tổng hợp, phân tích và báo cáo. Chỉ nghiên cứu sinh và giám sát viên được tiếp nhận những thông tin này. Chúng tôi luôn cố gắng hết sức để đảm bảo bảo mật thông tin trong phạm vi cho phép của pháp luật. Cô/ bác/ anh/ chị sẽ được xem lại và chỉnh sửa các bản tường thuật trước khi số liệu được đưa đi phân tích.

**Lưu trữ số liệu:** Thông tin chưa qua xử lý, bao gồm băng ghi âm và bản tường thuật sẽ được cất giữ bằng khoá và mật mã, trong máy tính ở XXX trong vòng ít nhất 10 năm kể từ khi báo cáo kết quả đầu tiên được xuất bản. Thông tin đã qua mã hoá sẽ được chia sẻ cho mục đích nghiên cứu lâu dài.

**Quyền tiếp cận thông tin & liên lạc:** Nếu có thắc mắc gì, cô/ bác/ anh/ chị có thể hỏi bất cứ lúc nào. Nếu có thắc mắc trong quá trình sau này hoặc muốn rút khỏi đề án, vui lòng liên hệ nghiên cứu sinh Li Nguyen, số ĐT 0432 xxx xxx, hoặc e-mail nhxxx@cam.ac.uk. Nếu cô/ bác/ anh/ chị cần tư vấn tâm lí trong suốt quá trình thực hiện dự án, xin vui lòng liên hệ đường dây nóng Lifeline Help line, số ĐT 13 11 14.

**Đồng ý tự nguyện:** Nếu vui lòng hợp tác, cô/ bác/ anh/ chị vui lòng kí bên dưới. Chữ kí cũng đồng nghĩa với việc xác nhận thông tin đã được trao đổi kĩ càng và mọi thắc mắc đã được giải toả đầy đủ. Cô/ bác/ anh/ chị có thể tiếp tục phản hồi thắc mắc trong suốt quá trình nghiên cứu

hoặc sau đó nữa. Chú kí cũng xác nhận sự đồng ý tự nguyện tham gia dự án này và cho phép sử dụng, chia sẻ dữ liệu cho mục đích nghiên cứu.

Những vấn đề liên quan đến đạo đức của dự án này đã được Ủy ban Đạo Đức của ĐH Cambridge thông qua.

**Kí tên:**..... **Ngày:** .....

**Dưới 18 tuổi:**

**Chữ kí của người giám hộ:** ..... **Ngày:** .....

## C.2 English version

**Primary researcher:** Li Nguyen, Doctorate student in Linguistics, Faculty of Medieval and Modern Languages, University of Cambridge, UK.

**Project:** Vietnamese bilingual communication in Canberra

**Description and Methodology:** This research project investigates Vietnamese bilingual communication in Canberra.

**Participants:** First-generation migrant Vietnamese who has lived in Canberra for at least 10 years, or second generation Vietnamese who was either born in Canberra (and lived here since), or those who arrived by the year of 5. Speakers use both English and Vietnamese on a regular basis. Due to the limited focus and data transparency, participants who do not satisfy these conditions are not included. Data will be collected via a natural recorded talk and a questionnaire. The targets are adults over 18 years old, but if young children happen to be involved in the conversation, written consent to use their data will be sought from the child themselves and their caregivers.

**Use of data and feedback:** The results of the research will be published as a PhD thesis, and possibly disseminated through conference papers and/or journal articles. Participants can request a copy of the research output.

**What you will do in this project:** For this research, you will be asked to record a natural conversation of at least 30 minutes and fill in a questionnaire afterwards. The natural conversation is audio-taped and transcribed. The researcher is not present during the recording, you can decide when and what to tape within the timeframe given (expected within 2–3 weeks). The questionnaire can be emailed or posted to you, depending on your preference. Your total time commitment for this research is 50–60 minutes.

**Risk and confidentiality:** There are no obvious risks associated with this research. Its purpose is to observe your interactional linguistic behaviours in the community; the research does not intend to modify these in any way. Participation is voluntary, you can leave the research project at any point in time or decline to answer any questions without being asked to explain the reasons. If you decide to withdraw, data already collected from you will be destroyed and not used. If you still wish for your previously collected data to be used, please advise the researcher.

All personal identifiers will be removed from the material, including the audio-recordings before the data is collated, analysed and research outcomes produced. Only the primary researcher and the supervisor will have access to the raw data. Every necessary step will be taken to ensure confidentiality as far as the law allows. You will have an opportunity to review and edit your recording before giving it to the researcher.

**Data storage:** The raw data, including the transcripts and the questionnaire will be stored under a password protected Cambridge drive throughout the research and for a minimum of ten years following any publications arising from the research. Anonymised data however will be made readily available for research purposes.

**Right to ask questions and contact information:** If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact me on 0432 xxx xxx, or nhxxx@cam.ac.uk. If you feel distressed at any stage of the research, please contact Lifeline Help line on 13 11 14.

**Voluntary consent:** By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You understand that you may ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study and have your data used and shared for research purposes.

The ethical aspects of this research have been approved by the Cambridge Committee of Research Ethics in Linguistics.

**Signature:**..... **Date:** .....

**Under 18:**

**Signature of the guardian:** ..... **Date:** .....



## CANVEC CORPUS CONSTITUTION

This appendix presents detailed individual configurations of all recordings in CanVEC (Chapter 3, §3.2.1).

File name	Duration	N. speakers	Generation	Gender
Billy.Tyler.Ellie.0807	00:31:11	3	1.2.1	MMF
Brian.Dany.0812	00:25:35	2	1.1	MM
Hannah.Lida.0718	00:32:17	2	2.2	FF
Harry.Tressie.Josh.0719	00:29:43	3	1.2.2	MFm
Heather.Troy.0708	00:23:25	2	1.2	Fm
Helen.Vivian.Quinn.0818	00:21:50	3	1.2.2	FFm
Lami.Dany.0825	00:34:10	2	2.1	FF
Lina.Naomi.0623	00:15:46	2	1.2	FF
Luna.Tressie.0901	00:15:12	2	1.2	FF
Max.Thomas.0823	00:29:02	2	1.1	MM
Mia.Phoebe.0905	00:13:46	2	1.1	FF
Mina.Pete.0906	00:32:34	2	1.2	Fm
Penny.Marie.Rory.0912	00:24:37	3	2.1.1	FFm
Quentin.Sony.0306	00:26:38	2	1.1	MM
Quintus.Daniel.0711	00:24:53	2	1.1	MM
Reece.Taylor.0906	00:45:07	2	1.2	Mf
Tanner.Nina.0609	00:16:09	2	1.2	Mf
Tee.Taz.0808	00:23:00	2	1.1	Mf
Theresa.Twee.0715	00:28:47	2	1.2	FF
Tim.Jess.0629	00:13:08	2	1.2	Mf
Tim.Jess.Chloe.0705	00:18:36	3	1.2.1	MFF
Tom.Henry.0725	00:36:47	2	1.2	MM
Tom.Henry.0809	00:40:37	2	1.2	MM

Table D.1: Details of each recording in CanVEC, in alphabetical order. Information of speakers' generation and gender is recorded in order of the participants displayed in the file name.



---

## QUESTIONNAIRE

---

This appendix presents the bilingual questionnaire, which was used to collect participants' demographic and linguistic information (Chapter 3, §3.2.2).

### E.1 Vietnamese version

1. Giới tính:

Nam

Nữ

2. Ngày tháng năm sinh: .....

3. Nghề nghiệp: .....

4. Bằng cấp cao nhất hiện tại:

Tiểu học

Lớp 9

Lớp 12

Đại học & trên đại học

5. Bạn bắt đầu học Tiếng Việt khi nào?

Từ lúc 2 tuổi hoặc nhỏ hơn

Từ lúc 4 tuổi hoặc nhỏ hơn

Từ tiểu học

- Từ trung học
  - Trưởng thành mới được học
6. Bạn bắt đầu học Tiếng Anh khi nào?
- Từ lúc 2 tuổi hoặc nhỏ hơn
  - Từ lúc 4 tuổi hoặc nhỏ hơn
  - Từ tiểu học
  - Từ trung học
  - Trưởng thành mới được học
7. Trên thang điểm từ 1 đến 4, bạn đánh giá tiếng Việt của mình ở mức nào?
- 1. Vài từ ngữ cơ bản
  - 2. Tự tin hội thoại cơ bản
  - 3. Khá tự tin hội thoại phức tạp
  - 4. Tự tin hội thoại phức tạp
8. Trên thang điểm từ 1 đến 4, bạn đánh giá tiếng Anh của mình ở mức nào?
- 1. Vài từ ngữ cơ bản
  - 2. Tự tin hội thoại cơ bản
  - 3. Khá tự tin hội thoại phức tạp
  - 4. Tự tin hội thoại phức tạp
9. Mẹ bạn (hoặc cô, dì, vú nuôi, bà) thường giao tiếp với bạn bằng ngôn ngữ nào khi bạn nhỏ?
- Tiếng Việt
  - Tiếng Anh
  - Tiếng Việt & Tiếng Anh
  - Khác (Nêu rõ).....
  - Không phù hợp thực tế
10. Cha bạn (hoặc chú, bác, ông) thường giao tiếp với bạn bằng ngôn ngữ nào khi bạn nhỏ?
- Tiếng Việt

- Tiếng Anh
- Tiếng Việt & Tiếng Anh
- Khác (Nêu rõ).....
- Không phù hợp thực tế

11. Bạn học tiểu học bằng ngôn ngữ nào?

- Tiếng Việt
- Tiếng Anh
- Tiếng Việt & Tiếng Anh
- Khác (Nêu rõ).....

12. Các năm sau tiểu học bạn đi học bằng ngôn ngữ nào?

- Tiếng Việt
- Tiếng Anh
- Tiếng Việt & Tiếng Anh
- Khác (Nêu rõ).....

13. Kể ra 5 người bạn nói chuyện thường xuyên hàng ngày, trực tiếp hoặc qua điện thoại, ví dụ: đối tác, con, bạn, đồng nghiệp, vâng vâng. Sau đó ghi lại ngôn ngữ từng người dùng để nói chuyện với bạn, như ví dụ sau đây:

Tên người hoặc tên quan hệ (Dùng tên giả nếu muốn)	Ngôn ngữ bạn dùng để nói chuyện thường xuyên với từng người: (đánh dấu tick vào từng cột cho từng hàng)			
	Tiếng Việt	Tiếng Anh	Anh Việt bằng nhau	Ngôn ngữ khác
1. Minh	✓			
2. Mẹ		✓		
3. Sếp			✓	
4. Mai				✓
5. Chị		✓		

Điền vào khung sau:

Tên người hoặc tên quan hệ (Dùng tên giả nếu muốn)	Ngôn ngữ bạn dùng để nói chuyện thường xuyên với từng người: (đánh dấu tick vào từng cột cho từng hàng)			
	Tiếng Việt	Tiếng Anh	Anh Việt bằng nhau	Ngôn ngữ khác
1.				
2.				
3.				
4.				
5.				

14. Trên thang điểm từ 1 đến 5, bạn đánh giá tiếng Việt ở thang điểm nào cho mỗi đặc tính sau đây? Khoanh tròn 1 số cho mỗi hàng.

	←—————→					
không thân thiện	1	2	3	4	5	thân thiện
không truyền cảm hứng	1	2	3	4	5	truyền cảm hứng
không có ích	1	2	3	4	5	có ích
xấu xí	1	2	3	4	5	hoàn mỹ

15. Trên thang điểm từ 1 đến 5, bạn đánh giá tiếng Anh ở thang điểm nào cho mỗi đặc tính sau đây? Khoanh tròn 1 số cho mỗi hàng.

	←—————→					
không thân thiện	1	2	3	4	5	thân thiện
không truyền cảm hứng	1	2	3	4	5	truyền cảm hứng
không có ích	1	2	3	4	5	có ích
xấu xí	1	2	3	4	5	hoàn mỹ

16. Bạn tự nhận mình là.....?

- Người Việt
- Người Úc
- Khác (nêu rõ):.....

17. Quan điểm của bạn về câu nói sau: “Thường ngày tôi giữ tiếng Anh và tiếng Việt riêng biệt, không lẫn lộn.”

- 1. Rất không đồng ý
- 2. Không đồng ý
- 3. Không có quan điểm cụ thể
- 2. Đồng ý
- 3. Rất đồng ý

18. Quan điểm của bạn về câu nói sau: “Trong cùng một đoạn hội thoại chúng ta không nên dùng Anh Việt lẫn lộn.”

- 1. Rất không đồng ý
- 2. Không đồng ý
- 3. Không có quan điểm cụ thể
- 2. Đồng ý
- 3. Rất đồng ý

## E.2 English version

1. Are you:

Male

Female

2. What is your date of birth? .....

3. What is your occupation? .....

4. What is your highest level of education?

Primary school

Year 9

Year 12

University degree or above

5. When did you first learn Vietnamese?

Since I was 2 years old or younger

Since I was 4 years old or younger

Since primary school

Since secondary school

I learned Vietnamese as an adult

6. When did you first learn English?

Since I was 2 years old or younger

Since I was 4 years old or younger

Since primary school

Since secondary school

I learned English as an adult

7. On a scale of 1 to 4, how well do you feel you can speak Vietnamese?

1. Only know some words and expressions

2. Confident in basic conversations
3. Fairly confident in extended conversations
4. Confident in extended conversations
8. On a scale of 1 to 4, how well do you feel you can speak English?
1. Only know some words and expressions
2. Confident in basic conversations
3. Fairly confident in extended conversations
4. Confident in extended conversations
9. Which language(s) did your mother (or a female caregiver) speak to you while you were growing up?
- Vietnamese
- English
- Vietnamese & English
- Other (Please specify).....
- Not applicable
10. Which language(s) did your father (or a male caregiver) speak to you while you were growing up?
- Vietnamese
- English
- Vietnamese & English
- Other (Please specify).....
- Not applicable
11. Through which language(s) were you predominantly taught at primary school?
- Vietnamese
- English
- Vietnamese & English
- Other (Please specify).....

12. Through which language(s) were you predominantly taught during your later school years?

- Vietnamese
- English
- Vietnamese & English
- Other (Please specify).....

13. Make a list below of five of the people you speak to most in your everyday life, either in person or on the phone, e.g. your partner, your child, a friend, a workmate etc. Then note which language(s) that person uses to speak with you, as shown in the sample table.

Name of person, or relationship (use fictitious names if you prefer)	Language mostly spoken with that person: (place a tick in one cell below for each line)			
	Vietnamese	English	Equally Vietnamese & English	Another language
1. Minh	✓			
2. Mother		✓		
3. Boss			✓	
4. Mai				✓
5. Sister		✓		

Please fill in the table below:

Name of person, or relationship (use fictitious names if you prefer)	Language mostly spoken with that person: (place a tick in one cell below for each line)			
	Vietnamese	English	Equally Vietnamese & English	Another language
1.				
2.				
3.				
4.				
5.				

14. How would you rate the Vietnamese language on a scale of 1 to 5 regarding the following properties? Circle one number in each line.

	←-----→	
unfriendly	1      2      3      4      5	friendly
uninspiring	1      2      3      4      5	inspiring
useless	1      2      3      4      5	useful
ugly	1      2      3      4      5	beautiful

15. How would you rate the English language on a scale of 1 to 5 regarding the following properties? Circle one number in each line.

	←-----→	
unfriendly	1      2      3      4      5	friendly
uninspiring	1      2      3      4      5	inspiring
useless	1      2      3      4      5	useful
ugly	1      2      3      4      5	beautiful

16. Do you consider yourself to be mainly.....?

- Vietnamese
- Australian
- Other (please specify):.....

17. To what extent do you agree with the following statement: “In everyday conversation, I keep the Vietnamese and English languages separate.”

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

18. To what extent do you agree with the following statement: “People should avoid mixing Vietnamese and English in the same conversation.”

1. Strongly disagree

- 2. Disagree
- 3. Neither agree nor disagree
- 4. Agree
- 5. Strongly agree

## CANVEC ANNOTATION CONVENTIONS

This appendix lists all the annotation conventions used in CanVEC (Chapter 3, §3.3).

Code	Meaning	CanVEC example
.	final Intonation Unit with a falling pitch	we just stare at each other.
?	final Intonation Unit with a rising pitch	so they don't have to wait to go into the cubicle?
,	continuing Intonation Unit	<i>nó tới cái stage,</i>
X	unclear speech	<b>X</b> ,
<X>	unclear syllable	<i>điện-thoại rằng hẳn &lt;X&gt;.</i>
<abc>	unclear syllable/speech; 'abc' represents the transcriber's best guess at content	on my <b>&lt;Explore page&gt;</b> on Instagram I just saw these boys,
<E>	unclear syllable; transcriber's best guess that the syllable is in English	today when we were doing mental <E>,
<V>	unclear syllable; transcriber's best guess that the syllable is in Vietnamese	<i>mà hình-sự nó cho prosecutor điều-tra là &lt;V&gt; ấy,</i>
[A:]	anonymised information, which includes person names and place names	then in summer she left to <i>đi</i> <b>[A:school name]</b> or something,

Table F.1: CanVEC transcription conventions (modelled on Du Bois et al., 1993)

Code	Meaning
@non	language-neutral token/ clause
@vie	Vietnamese token/ clause
@eng	English token/ clause
@mix	clause that contains tokens from both Vietnamese and English

Table F.2: CanVEC semi-automatic language tags



---

## VIETNAMESE TO UNIVERSAL POS TAG MAP

---

This appendix presents the defined tag map converting the Underthesea Tagset to the Universal Tagset to ensure consistencies of POS tags across CanVEC ([Chapter 3, §3.3.2.1](#)).

<b>Underthesea</b>	<b>Universal</b>	<b>Underthesea</b>	<b>Universal</b>
A	ADJ	Nc	CLS
ADP	PREP	Nu	NOUN
C	CONJ	Ny	PROPN
CCONJ	CONJ	P	PRON
E	PREP	R	ADV
I	INTJ	T	VERB
L	DET	V	VERB
M	NUM	X	X
N	NOUN	Z	Z

Table G.1: The mapping function for Underthesea POS tags to Universal POS tags. The CLS tag (classifiers) is not a universal tag, but was considered important to preserve for Vietnamese.



---

## VIETNAMESE POS-TAG CONFUSION MATRICES

---

This appendix reports the full confusion matrices of Vietnamese POS tags in the CanVEC evaluation sample (Chapter 3, §3.3.3).

### H.1 Sample Vietnamese clauses

Correct tag	Tagged as	N	%
Pronoun (PRON)	Classifier (CLS)	43	34.4%
	Noun (NOUN)	15	12.0%
	Particle (PRT)	7	5.6%
	Preposition (PREP)	2	1.6%
Classifier (CLS)	Pronoun (PRON)	45	36.0%
	Interjection (INTJ)	6	4.8%
Noun (NOUN)	Verb (VERB)	5	4.0%
Verb (VERB)	Adverb (ADV)	1	0.8%
Adverb (ADV)	Verb (VERB)	1	0.8%
<b>TOTAL</b>		<b>125</b>	<b>100%</b>

Table H.1: Confusion matrices of Vietnamese POS tags (N=520 tags) in the Vietnamese evaluation sample (N=100 clauses)

## H.2 Sample mixed clauses

Correct tag	Tagged as	N	%
Pronoun (PRON)	Classifier (CLS)	16	28.6%
	Noun (NOUN)	11	19.6%
Classifier (CLS)	Pronoun (PRON)	11	19.6%
	Interjection (INTJ)	6	10.7%
Noun (NOUN)	Proper Noun (PROPN)	3	5.4%
	Particle (PRT)	2	3.6%
Verb (VERB)	Adverb (ADV)	3	5.4%
Preposition (PREP)	Noun (NOUN)	2	3.6%
	Particle (PRT)	1	1.8%
	Verb (VERB)	1	1.8%
<b>TOTAL</b>		<b>56</b>	<b>100%</b>

Table H.2: Confusion matrices of Vietnamese POS tags (N=224 tags) in the mixed-clause evaluation sample (N=100 clauses)

## CANVEC EXAMPLE OF AN ANNOTATED DIALOGUE

---

This appendix (starting from the next page) presents an extended example of an annotated dialogue from a recording in CanVEC (Tim.Jess.Chloe.0705). The example was chosen because it showcases a mixture of all clause types in a short span of conversation. Tim and Chloe are first-generation speakers, and Jess is a second-generation speaker in this transcript. (Chapter 3, §3.3.4).

### NOTES:

- (i) This example was not part of the evaluation sample in Chapter 3, §3.3.3;
- (ii) It contains two automatic POS tagging errors and one translation error, all in the third clause (Tim, 00:07.4–00:13.9). Specifically,
  - *ba* was tagged as NUM, but should have been PRON;
  - *đi* was tagged as VERB, but should have been PRT; and
  - the translation was fluent and comprehensible but not semantically adequate; a better translation is provided in brackets for the reader;
- (iii) As a reminder from the main text, **@non** represents a language-neutral token, **@vie** represents a Vietnamese token/clause, **@eng** represents an English token/clause, and **@mix** represents a code-switching clause.

Speaker	Level	Annotation
Tim 00:04.3–00:05.7 <b>@vie</b>	Clause	<i>con có thích không?</i>
	Tokens	<i>con có thích không</i>
	POS	PRON VERB VERB ADV
	TokenLang	@vie @vie @vie @vie
	Translation	Do you like it?
Jess 00:05.7–00:07.4 <b>@vie</b>	Clause	<i>uhm thích.</i>
	Tokens	<i>uhm thích</i>
	POS	INTJ VERB
	TokenLang	@non @vie
	Translation	I like it.
Tim 00:07.4–00:13.9 <b>@mix</b>	Clause	<i>con kể cho ba nghe vài cái về cái concert đi.</i>
	Tokens	<i>con kể cho ba nghe vài cái về cái concert đi</i>
	POS	PRON VERB PREP NUM VERB NUM CLS PREP CLS NOUN VERB
	TokenLang	@vie @eng @vie
	Translation	*I told you something about the concert (Can you tell me something about the concert?)
Jess 00:13.9–00:20.3 <b>@eng</b>	Clause	<i>well the concert it has the Kpop boy band,</i>
	Tokens	<i>well the concert it has the Kpop boy band</i>
	POS	INTJ DET NOUN PRON VERB DET PROPN NOUN NOUN
	TokenLang	@eng @eng @eng @eng @eng @eng @non @eng @eng
	Translation	—

Speaker	Level	Annotation
Jess 00:20.3–00:27.7 <b>@mix</b>	Clause	BTS and like <i>có bảy cái</i> members.
	Tokens	BTS and like <i>có bảy cái</i> members
	POS	PROPN CONJ PREP VERB NUM CLS NOUN
	TokenLang	@non @eng @eng @vie @vie @vie @eng
	Translation	BTS and like has seven members.
Jess 00:27.7–00:30.8 <b>@eng</b>	Clause	and they are all really good looking.
	Tokens	and they are all really good looking
	POS	CONJ PRON VERB ADV ADV ADJ NOUN
	TokenLang	@eng @eng @eng @eng @eng @eng @eng
	Translation	—
Chloe 00:30.8–00:34.5 <b>@eng</b>	Clause	that is the main thing.
	Tokens	that is the main thing
	POS	DET VERB DET ADJ NOUN
	TokenLang	@eng @eng @eng @eng @eng
	Translation	—
Jess 00:34.5–00:37.5 <b>@eng</b>	Clause	no they are all really talented too.
	Tokens	no they are all really talented too
	POS	INTJ PRON VERB ADV ADV ADJ ADV
	TokenLang	@eng @eng @eng @eng @eng @eng @eng
	Translation	—

Speaker	Level	Annotation
Tim 00:37.5–00:45.0 <b>@vie</b>	Clause	<i>yeah rồi trong những cái bảy cái người đó thì con thích ai nhất và tại sao?</i>
	Tokens	<i>yeah rồi trong những cái bảy cái người đó thì con thích ai nhất và tại_sao</i>
	POS	INTJ PART PREP DET CLS NUM CLS NOUN DET CONJ PRON VERB PRON ADJ CONJ X
	TokenLang	@vie
	Translation	yeah then in those seven people who do you like best and why?
Jess 00:45.0–00:48.5 <b>@vie</b>	Clause	<i>con thích Jimmy nhất,</i>
	Tokens	<i>con thích Jimmy nhất</i>
	POS	PRON VERB PROPN NUM
	TokenLang	@vie @vie @non @vie
	Translation	I like Jimmy the most,
Jess 00:48.5–00:54.3 <b>@mix</b>	Clause	<i>tại vì he is my ideal type.</i>
	Tokens	<i>tại vì he is my ideal type</i>
	POS	PREP PREP PRON VERB ADJ ADJ NOUN
	TokenLang	@vie @vie @eng @eng @eng @eng @eng
	Translation	because he is my ideal type.
Tim 00:54.3–00:56.7 <b>@eng</b>	Clause	<i>what is your ideal type?</i>
	Tokens	<i>what is your ideal type</i>
	POS	NOUN VERB ADJ ADJ NOUN
	TokenLang	@eng @eng @eng @eng @eng
	Translation	—

Speaker	Level	Annotation
Jess 00:56.7–01:00.1 <b>@eng</b>	Clause	boyish good look,
	Tokens	boyish good look
	POS	ADJ ADJ NOUN
	TokenLang	@eng @eng @eng
	Translation	—
Jess 01:00.1–01:04.7 <b>@eng</b>	Clause	you know like Leonardo DiCaprio when he was younger.
	Tokens	you know like Leonardo DiCaprio when he was younger
	POS	PRON VERB PREP PROPN PROPN ADV PRON VERB ADJ
	TokenLang	@eng @eng @eng @non @non @eng @eng @eng @eng
	Translation	—
Jess 01:04.7–01:07.6 <b>@eng</b>	Clause	not like too manly.
	Tokens	not like too manly
	POS	ADV PREP ADV ADJ
	TokenLang	@eng @eng @eng @eng
	Translation	—



## CANVEC VIETNAMESE INTRANSITIVE VERBS

This appendix lists all the Vietnamese intransitive verbs in CanVEC, which are straightforwardly excluded in the analysis of null direct objects (Chapter 5, §5.5.1.2).

Vietnamese Verb	English Translation	Vietnamese Verb	English Translation
<i>biển</i>	go away	<i>lăn</i>	roll off
<i>bơi</i>	swim	<i>mơ</i>	dream
<i>bước</i>	step	<i>mưa</i>	rain
<i>bể</i>	broken	<i>nghĩ</i>	think
<i>chạy</i>	run	<i>nghỉ</i>	rest
<i>chảy</i>	flow	<i>ngã</i>	fall
<i>chết</i>	die	<i>ngồi</i>	sit
<i>cố-gắng</i>	try (to do something)	<i>ngủ</i>	sleep
<i>du-học</i>	study abroad	<i>nhảy</i>	dance
<i>du-lịch</i>	travel	<i>nổ</i>	explode
<i>dạo</i>	stroll	<i>quỳ</i>	kneel down
<i>dậy</i>	get up	<i>ra-đời</i>	come to life
<i>đau</i>	hurt*	<i>reng</i>	ring*
<i>đi</i>	go	<i>rơi</i>	fall
<i>đi-lại</i>	travel	<i>rẽ</i>	turn*
<i>đứng</i>	stand	<i>sập</i>	collapse
<i>đứt</i>	broken	<i>sống</i>	live
<i>giải-lao</i>	take a break	<i>tan</i>	melt*
<i>gãy</i>	broken	<i>thành-công</i>	succeed
<i>hành-động</i>	act	<i>té</i>	fall
<i>khóc</i>	cry	<i>tập-trung</i>	focus
<i>lui</i>	back off	<i>vỡ</i>	broken
<i>lên-tiếng</i>	speak up	<i>xuất-hiện</i>	appear

Table J.1: Vietnamese intransitive verbs in CanVEC. Asterisks (\*) denote words that are optionally transitive in English, but exclusively intransitive in Vietnamese.



THE DISTRIBUTION OF VIETNAMESE NULL  
SUBJECTS PER SPEAKER

---

This appendix (starting from the next page) provides a breakdown of null subjects produced per grammatical person per speaker ([Chapter 5, §5.7.1.2](#)).

## K.1 First-generation speakers

Speaker	First person		Second person		Third person		Total
	N	%	N	%	N	%	
Tee	47	95%	3	5%	0	0%	50
Rory	23	90%	1	3%	2	7%	26
Reece	33	89%	3	8%	1	3%	37
Thomas	32	89%	3	8%	1	3%	36
Luna	29	88%	0	0%	4	12%	33
Quintus	39	83%	5	11%	3	6%	47
Max	36	82%	2	5%	6	13%	44
Mia	31	80%	3	7%	5	13%	39
Tim	54	79%	5	7%	9	14%	68
Tom	38	78%	5	10%	6	12%	49
Marie	44	77%	5	9%	8	14%	57
Mina	39	75%	6	12%	7	13%	52
Taz	49	71%	6	9%	14	20%	69
Phoebe	22	71%	4	13%	5	16%	31
Ellie	33	69%	5	10%	10	21%	48
Harry	34	65%	3	6%	15	29%	52
Daniel	34	65%	10	19%	8	16%	52
Sony	29	60%	6	13%	13	27%	48
Chloe	22	60%	7	18%	8	22%	37
Dany	26	58%	9	20%	10	22%	45
Heather	30	57%	12	22%	11	21%	53
Theresa	31	55%	7	13%	18	32%	56
Quentin	32	53%	13	22%	15	25%	60
Helen	24	50%	12	25%	12	25%	48
Lina	31	48%	15	23%	19	29%	65
Brian	11	48%	5	22%	7	30%	23
Tanner	16	47%	7	21%	11	32%	34
Billy	24	45%	14	27%	15	28%	53
<b>TOTAL</b>							<b>1311</b>

Table K.1: First generation speakers' numerical distribution of Vietnamese null subjects by grammatical person

## K.2 Second-generation speakers

Speaker	First person		Second person		Third person		Total
	N	%	N	%	N	%	
Tressie	1	6%	14	82%	2	12%	17
Twee	1	4%	19	80%	4	16%	24
Troy	1	6%	12	80%	2	14%	15
Taylor	3	19%	12	75%	1	6%	16
Quinn	2	13%	10	67%	3	20%	15
Pete	2	17%	8	66%	2	17%	12
Lami	5	31%	10	63%	1	6%	16
Josh	1	12%	5	63%	2	25%	8
Vivian	3	17%	11	61%	4	22%	18
Tyler	3	20%	9	60%	3	20%	15
Jess	3	20%	9	60%	3	20%	15
Penny	2	12%	10	59%	5	29%	17
Naomi	2	22%	5	56%	2	22%	9
Henry	3	23%	7	54%	3	23%	13
Nina	5	26%	10	53%	4	21%	19
Lida	6	35%	5	30%	6	35%	17
Hannah	3	25%	4	33%	5	42%	12
<b>TOTAL</b>							<b>258</b>

Table K.2: Second-generation speakers' numerical distribution of Vietnamese null subjects by grammatical person



## REFERENCES

---

- Aalberse, Suzanne, Ad Backus & Pieter Muysken. 2019. *Heritage languages: A language contact approach*. Amsterdam: John Benjamins Publishing Company. doi:10.1075/sibil.58. Cited on pages 1, 172, and 206.
- Aaron, Jessi Elana. 2015. Lone English-origin nouns in Spanish: The precedence of community norms. *International Journal of Bilingualism* 19(4), 459–480. doi:10.1177/1367006913516021. Cited on page 51.
- Adamou, Evangelia, Stefano De Pascale, Yekaterina García-Márkina & Cristian Padure. 2019. Do bilinguals generalize *estar* more than monolinguals and what is the role of conceptual transfer? *International Journal of Bilingualism* 23(6), 1549–1580. doi:10.1177/1367006918812175. Cited on page 154.
- Adamou, Evangelia De Pascale. 2013. Replicating Spanish *estar* in Mexican Romani. *Linguistics* 51(6), 1075–1105. Cited on page 154.
- Adams, Anne & Anna L. Cox. 2008. Questionnaires, in-depth interviews and focus groups. In Paul Cairns & Anna L. Cox (eds.), *Research methods for human computer interaction*, 17–34. Cambridge: Cambridge University Press. <http://oro.open.ac.uk/11909/>. Cited on page 36.
- Aikhenvald, Alexandra Y. 2002. *Language contact in Amazonia*. Oxford: Oxford University Press. Cited on page 82.
- Alba, Richard, John Logan, Amy Lutz & Brian Stults. 2002. Only English by the third generation? Loss and preservation of the mother tongue among the grandchildren of contemporary immigrants. *Demography* 39(3), 467–484. Cited on page 69.
- Allerton, D.J. 1975. Deletion and pro-form reduction. *Journal of Linguistics* 11, 213–238. Cited on page 128.
- Allerton, D.J. 1982. *Valency and the English verb*. London: Academic Press. Cited on page 128.

- Alonso-Ovalle, Luis, Susana Fernández-Solera, Lyn Frazier & Jr Charles Clifton. 2002. Null vs. Overt pronouns and the topic-focus articulation in Spanish. *Italian Journal of Linguistics* 14, 151–170. Cited on page 187.
- Auer, Peter. 2002. Introduction. In Peter Auer (ed.), *Style and social identities: Alternative approaches to linguistic heterogeneity*, 1–21. Berlin: de Gruyter. Cited on page 123.
- Auer, Peter & Raihan Muhamedova. 2005. ‘Embedded language’ and ‘matrix language’ in insertional language mixing: Some problematic cases. *Italian Journal of Linguistics* 17(1), 35–54. Cited on pages 71, 74, 79, and 99.
- Australian Bureau of Statistics. 2017. Census 2016: Australian Capital Territory. Tech. rep. ABS Canberra, ACT. Cited on pages 2, 13, 16, and 115.
- Backus, Ad. 2005. Codeswitching and language change: One thing leads to another? *International Journal of Bilingualism* 9(3-4), 307–340. Cited on pages 132 and 177.
- Bailey, Beryl. 1966. *Jamaican creole syntax*. Cambridge: Cambridge University Press. Cited on page 134.
- Bailey, Guy & Jan Tillery. 2004. Some sources of divergent data in sociolinguistics. In Ronald Macaulay & Carmen Fought (eds.), *Sociolinguistic variation: Critical reflections*, 11–30. Oxford: Oxford University Press. Cited on page 146.
- Bakker, Peter (ed.). 1997. *A language of our own: The genesis of Michif, the mixed Cree-French Language of the Canadian Métis*. Oxford University Press. Cited on page 71.
- Balukas, Colleen & Christian Koops. 2015. Spanish-English bilingual voice onset time in spontaneous code-switching. *International Journal of Bilingual Education and Bilingualism* 19(4), 423–443. doi:10.1177/1367006913516035. Cited on page 53.
- Barbosa, Pilar, Maria Eugênia L. Duarte & Mary Aizawa Kato. 2005. Null subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics* 4(2), 11–52. doi:10.5334/jpl.158. Cited on page 147.
- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48. doi:10.18637/jss.v067.i01. Cited on page 157.
- Baugh, John. 1980. A reexamination of the Black English copula. In William Labov (ed.), *Locating Language in Time and Space*, 83–106. New York: Academic Press. Cited on page 154.

- Bayley, Robert & Lucinda Pease-Alvarez. 1997. Null pronoun variation in Mexican-descent children's narrative discourse. *Language Variation and Change* 9(3), 349–371. Cited on page 147.
- Becker, Kara. 2014. (r) we there yet? The change to rhoticity in New York City English. *Language Variation and Change* 26(2), 141–168. doi:10.1017/S0954394514000064. Cited on page 122.
- Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2), 145–204. Cited on page 122.
- Bell, Allan. 1999. Styling the other to define the self: A study in New Zealand identity making. *Journal of Sociolinguistics* 3(4), 523–541. doi:10.1111/1467-9481.00094. Cited on page 122.
- Ben-Mosche, Danny & Joanne Pyke. 2012. The Vietnamese diaspora in Australia: Current and potential links with the homeland. Report of an Australian Research Council linkage project. Tech. rep. The Australian Research Council. Cited on pages 13, 14, 15, 29, and 31.
- Benmamoun, Elabbas, Silvina Montrul & Maria Polinsky. 2013. Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics* 39(3-4), 129–181. <https://www.degruyter.com/view/journals/thli/39/3-4/article-p129.xml>. Cited on page 181.
- Bentahila, Abdelâli & Eirlys E. Davies. 1983. The syntax of Arabic-French code-switching. *Lingua* 59, 301–330. Cited on page 99.
- Berk-Seligson, Susan. 1986. Linguistic constraints on intra-sentential code-switching: A study of Spanish/Hebrew bilingualism. *Language in Society* 15, 313–348. Cited on page 99.
- Betts, Katharine. 2001. Boat people and public opinion in Australia. *People and Place* 9(4), 34–48. Cited on page 12.
- Biberauer, Theresa. 2017. Factors 2 and 3: A principled approach. In Chenchen Song & James Baker (eds.), *Cambridge Occasional Papers in Linguistics*, vol. 10, 38–65. Cambridge: Theoretical and Applied Linguistics, University of Cambridge. Cited on pages 114 and 180.
- Biberauer, Theresa. 2018. Less is more: Some thoughts on the tolerance principle in the context of the maximise minimal means model. In Chenchen Song, Li Nguyen & James Baker (eds.), *Cambridge Occasional Papers in Linguistics, Vol. 11*, 131–145. Cambridge: Theoretical and Applied Linguistics, University of Cambridge. Cited on page 180.
- Biberauer, Theresa. 2019. Children always go beyond the input: The maximise minimal means perspective. *Theoretical Linguistics* 45(3-4), 211–224. doi:10.1515/tl-2019-0013. Cited on pages 114 and 180.

- Biberauer, Theresa, Anders Holmberg, Ian Roberts & Michelle Sheehan. 2010. *Parametric variation null subjects in minimalist theory*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511770784. Cited on page 124.
- Birdsong, David. 2004. Second language acquisition and ultimate attainment. In *The handbook of applied linguistics*, 82–105. Hoboken, NJ: John Wiley & Sons, Ltd. doi:10.1002/9780470757000.ch3. Cited on page 116.
- Blommaert, Jan. 2011. Pragmatics and discourse. In Rajend Mesthrie (ed.), *The cambridge handbook of sociolinguistics*, 122–137. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511997068.012. Cited on page 183.
- Bolonyai, Agnes. 2000. "Elective affinities": Language contact in the abstract lexicon and its structural consequences: 'Affinities really become interesting only when they bring about separations' (Goethe: Elective affinities). *International Journal of Bilingualism* 4(1), 81–106. doi:10.1177/13670069000040010601. Cited on pages 132 and 177.
- Bolton, Kingsley. 2018. World Englishes and second language acquisition. *Special Issue: World Englishes and Second Language Acquisition* 37(1), 5–18. Cited on page 78.
- Bouchard, Marie-Eve. 2018. Subject pronoun expression in Santomean Portuguese. *Journal of Portuguese Linguistics* 17(1), 5. doi:http://doi.org/10.5334/jpl.191. Cited on page 147.
- Boussofara-Omar, Naima. 2003. Revisiting Arabic diglossic switching in light of the MLF model and its sub-models: The 4-M model and the Abstract Level model. *Bilingualism: Language and Cognition* 6(1), 33–46. Cited on pages 71, 72, 73, and 99.
- Brazil, David. 1985. *The communicative value of intonation in English*. Discourse analysis monograph 8. Birmingham: Bleak House : English Language Research. Cited on page 41.
- Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press. Cited on page 115.
- Brunelle, Marc & Xuyen Le. 2014. Why is sound symbolism so common in Vietnamese? In Jeffrey P. Williams (ed.), *The aesthetics of grammar: Sound and meaning in the languages of mainland Southeast Asia*, 83–98. Cambridge: Cambridge University Press. Cited on pages 48, 124, and 127.
- Bucholtz, Mary. 1999. "Why be normal?" Language and identity practices in a community of nerd girls. *Language in Society* 28, 203–223. Cited on page 122.

- Buffington, Albert & Preston Barba. 1954. *A Pennsylvania German grammar*. Allentown, PA: Schlechter's. Cited on page 80.
- Bullock, Barbara, Jacqueline Serigos, Almeida Jacqueline Toribio & Arthur Wendorf. 2018a. Predicting the presence of a matrix language in code-switching. In *CodeSwitch@ACL*, . Cited on page 48.
- Bullock, Barbara, Jacqueline Serigos, Almeida Jacqueline Toribio & Arthur Wendorf. 2018b. The challenges and benefits of annotating oral bilingual corpora. *Linguistic Variation* 18(1), 100–119. doi:10.1075/lv.00006.bul. Cited on pages 36 and 47.
- Bullock, Barbara & Almeida Jacqueline Toribio (eds.). 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press. doi:10.1017/CBO9780511576331. Cited on page 65.
- Burnley, Ian. 1989. Settlement dimensions of the Vietnam-born population in metropolitan Sydney. *Australian Geographical Studies* 27(2), 129–154. doi:10.1111/j.1467-8470.1989.tb00599.x. Cited on page 13.
- Butt, John & Carmen Benjamin. 2004. *A new reference grammar of modern Spanish*. London: Edward Arnold. Cited on page 149.
- Buttery, Paula, Michael McCarthy & Ronald Carter. 2015. Chatting in the academy: Informality in spoken academic discourse. In Nicholas Groom, Maggie Charles & Suganthi John (eds.), *Corpora, Grammar and Discourse*, 183–210. Amsterdam: John Benjamins Publishing Company. Cited on page 96.
- Caines, Andrew, Christian Bentz, Calbert Graham, Tim Polzehl & Paula Buttery. 2016. Crowdsourcing a multi-lingual speech corpus: Recording, transcription and annotation of the CROWDED CORPUS. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*, Portorož, Slovenia: European Language Resources Association (ELRA). Cited on pages 36 and 96.
- Caines, Andrew & Paula Buttery. 2010. You talking to me? A predictive model for zero-auxiliary constructions. In *2010 Workshop on NLP and linguistics: Finding the common ground, ACL-2010 Association for Computational Linguistics 2010, Uppsala, Sweden, July 11–16 proceedings*, 43–51. <http://www.aclweb.org/anthology/W10-2107>. Cited on page 96.
- Cantone, Francesca & Jeff MacSwan. 2009. Adjectives and word order: A focus on Italian-German code-switching. In Kees de Bot, Ludmila Isurin & Donald Winford (eds.), *Multi-*

- disciplinary approaches to code-switching*, 243–278. Amsterdam: John Benjamins Publishing Company. Cited on page 76.
- Cao, Xuan Hao. 2003. *Tiếng việt – mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa Vietnamese: Issues in phonetics, syntax, and semantics*. Ho Chi Minh City: Nhà Xuất bản Giáo Dục. Cited on pages 109 and 111.
- Carminati, Maria N. 2002. The processing of Italian subject pronoun. University of Massachusetts Amherst Phd dissertation. Cited on page 187.
- Carpenter, Kathie Lou. 1987. How children learn to classify nouns in Thai. Stanford University Phd dissertation. Cited on page 111.
- Carruthers, Ashley. 2008a. Vietnamese. In *The dictionary of Sydney*, Sydney: <https://dictionaryofsydney.org/entry/vietnamese>. Cited on pages 12, 13, and 46.
- Carruthers, Ashley. 2008b. Saigon from the diaspora. *Singapore Journal of Tropical Geography* 29(1), 68–86. doi:10.1111/j.1467-9493.2008.00320.x. Cited on page 15.
- Carter, Diana, Margaret Deuchar, Peredur Davies & María Del Carmen Parafita Couto. 2011. A systematic comparison of factors affecting the choice of matrix language in three bilingual communities. *Journal of Language Contact* 4(2). doi:10.1163/187740911X592808. Cited on page 96.
- Carter, Phillip M. & Tonya E. Wolford. 2018. Grammatical change in borderlands Spanish: A variationist analysis of copula variation and progressive expansion in a south texas bilingual enclave community. *Studies in Hispanic and Lusophone Linguistics* 11(1), 1–27. doi:10.1515/shll-2018-0001. Cited on pages 179 and 180.
- Carter, Ronald & Michael McCarthy. 2017. Spoken grammar: Where are we and where are we going? *Applied Linguistics* 38(1), 1–20. doi:10.1093/applin/amu080. Cited on pages 40 and 96.
- Čermáková, Anna, Zuzana Komrsková, Marie Kopřivová & Petra Poukarová. 2017. Between syntax and pragmatics: The causal conjunction protože in spoken and written Czech. *Corpus Pragmatics* 1(4), 393–414. doi:10.1007/s41701-017-0014-y. Cited on page 96.
- Chafe, Wallace. 1980. The deployment of consciousness in the production of a narrative. In Wallace Chafe (ed.), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*, Norwood, NJ: Ablex. Cited on page 41.

- Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, Amsterdam: John Benjamins Publishing Company. Cited on page 40.
- Chafe, Wallace. 1994. Intonation units. In Wallace Chafe (ed.), *Discourse, consciousness and time: The flow and displacement of conscious experience in speaking and writing*, Chicago: University of Chicago Press. Cited on pages 40, 41, 42, and 149.
- Chambers, Jack. 1995. *Sociolinguistic theory: Linguistic variation and its social significance*. Oxford: Blackwell. Cited on page 155.
- Champaud, Christian & Dominique Bassano. 1994. French concessive connectives and argumentation: In experimental study in eight- to ten-year-old children. *Journal of Child Language* 21(2), 415–438. doi:10.1017/S0305000900009338. Cited on page 34.
- Chan, Brian Hok-Shing. 2008. Code-switching, word order and the lexical/functional category distinction. *Lingua* 118(6), 777–809. Cited on page 99.
- Chan, Brian Hok-Shing. 2009. Code-switching with typologically distinct languages. In Barbara Bullock & Almeida Jacqueline Toribio (eds.), *Cambridge handbooks in linguistics. The Cambridge handbook of linguistic code-switching*, 182–198. Cambridge: Cambridge University Press. doi:10.1016/j.lingua.2007.05.004. Cited on pages 83 and 92.
- Cheshire, Jenny & Susan Fox. 2016. From sociolinguistic research to English language teaching. In Karen P. Corrigan & Adam Mearns (eds.), *Creating and digitizing language corpora: Volume 3: Databases for public engagement*, 265–290. London: Palgrave Macmillan. doi:10.1057/978-1-137-38645-8\_10. Cited on pages 45 and 96.
- Cheshire, Jenny, Paul Kerswill, Sue Fox & Eivind Torgersen. 2011. Contact, the feature pool and the speech community: The emergence of multicultural London English. *Journal of Sociolinguistics* 15(2), 151–196. doi:10.1111/j.1467-9841.2011.00478.x. Cited on page 38.
- Choi, Jinny. 2000. [–Person] direct object drop: The genetic cause of a syntactic feature in Paraguayan Spanish. *Hispania* 83(3), 531–543. doi:10.2307/346046. Cited on page 153.
- Chomsky, Noam. 1995. *The minimalist program*. Cambridge, MA: MIT Press. Cited on pages 175 and 181.
- Christensen, Matthew B. 2000. Anaphoric reference in spoken and written Chinese narrative discourse. *Journal of Chinese Linguistics* 28(2), 303–336. Cited on page 149.

- Cinque, Guglielmo. 1994. On the evidence for partial N-movement in the Romance DP. In Guglielmo Cinque, Jan Koster, Jean-Yves Pollock, Luigi Rizzi & Raffaella Zanuttini (eds.), *Paths towards Universal Grammar. Studies in honor of Richard S. Kayne*, 85–110. Washington, DC: Georgetown University Press. Cited on page 75.
- Clark, Eve. 2016. *First language acquisition*. Cambridge: Cambridge University Press. Cited on page 78.
- Clark, Marybeth. 1988. Vietnamese language and culture. In T.N. Nien Tran, H. Nguyen & L. Le (eds.), *Vietnamese language and attitudes towards personal relations*, 21–25. South Australia: Vietnamese Community in Australia. Cited on page 126.
- Clark, Marybeth. 1992. Conjunction as topicaliser in Vietnamese. *Mon-Khmer Studies: Special issue for Laurence Thompson's 60th birthday* 20, 91–109. Cited on page 129.
- Clark, Marybeth. 1996. Conjunction as copula in Vietnamese. *Mon-Khmer Studies* 26, 319–331. Cited on pages 129, 130, and 180.
- Clyne, Michael. 1987. Constraints on code-switching: How universal are they? *Linguistics* 25(4), 739–764. doi:10.1515/ling.1987.25.4.739. Cited on pages 83 and 99.
- Clyne, Michael. 2003. *Dynamics of language contact: English and immigrant languages*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511606526. Cited on pages 107, 135, 165, and 210.
- Colantoni, Laura. 2002. Clitic doubling, null objects and clitic climbing in the Spanish of corrientes. In Javier Gutiérrez-Rexach (ed.), *From words to discourse: Trends in Spanish semantics and pragmatics*, 321–336. Oxford: Elsevier. Cited on page 153.
- Cornips, Leonie & Karen Corrigan. 2005. Toward an integrated approach to syntactic variation a retrospective and prospective synopsis. In Leonie Cornips & Karen Corrigan (eds.), *Syntax and variation: Reconciling the biological and the social*, 1–27. Amsterdam: John Benjamins Publishing Company. Cited on page 6.
- Costa, Albert, Alfonso Caramazza & Nuria Sebastian Galles. 2000. The cognate facilitation effect: Implications for models of lexical access. *Journal of experimental psychology: Learning, memory, and cognition* 26(5), 1283–1296. doi:10.1037//0278-7393.26.5.1283. Cited on page 184.
- Coupland, Nikolas. 2007. *Style: Language variation and identity. Key topics in sociolinguistics*. Cambridge: Cambridge University Press. Cited on page 122.

- Coupland, Nikolas. 2009. Dialect style, social class and metacultural performance: The pantomime dame. In Rajend Mesthrie (ed.), *The new sociolinguistics reader*, 311–325. Basingstoke: Palgrave Macmillan. Cited on page 122.
- Coupland, Nikolas. 2011. The sociolinguistics of style. In Rajend Mesthrie (ed.), *The Cambridge handbook of Sociolinguistics*, 138–156. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511997068.013. Cited on page 122.
- Cournane, Ailís. 2019. A developmental view on incrementation in language change. *Theoretical Linguistics* 45(3-4), 127–150. doi:10.1515/tl-2019-0010. Cited on page 114.
- Cresti, Emanuela. 2014. Syntactic property of spontaneous speech in the language into act theory: Data on Italian complements and relative clauses. In Tommaso Raso & Heliana Mello (eds.), *Spoken corpora and Linguistic studies*, 365–410. Amsterdam: John Benjamins Publishing Company. Cited on page 96.
- Crible, Ludivine & Maria-Josep Cuenca. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse* 8(2), 149–166. doi:10.5087/dad.2017.207. Cited on page 96.
- Cutler, Cecilia A. 1999. Yorkville crossing: White teens, hip hop and African American English. *Journal of Sociolinguistics* 3(4), 428–442. doi:10.1111/1467-9481.00089. Cited on page 122.
- Cztinglar, Christine, Antigone Katicic, Katharina Kähler & Chris Schaner-Wolles. 2008. Strategies in the L1-acquisition of predication: The copula construction in German and Croatian. In Natalia Gagarina & Insa Gulzow (eds.), *The acquisition of verbs and their grammar: The effect of particular languages*, 71–104. Dordrecht: Springer. doi:10.1007/978-1-4020-4335-2\_4. Cited on page 155.
- Dal Negro, Silvia. 2004. Language contact and dying languages. *Revue Française de Linguistique Appliquée* 9(2), 47–58. doi:10.3917/rfla.092.0047. Cited on page 5.
- Dannenberg, Clare. 2002. Grammatical and phonological manifestations of null copula. *The Publication of the American Dialect Society* 87(1), 71–83. doi:10.1215/-87-1-71. Cited on page 134.
- de Prada Pérez, Ana. 2019. Theoretical implications of research on bilingual subject production: The vulnerability hypothesis. *International Journal of Bilingualism* 23(2), 670–694. doi:10.1177/1367006918763141. Cited on page 179.

- de Prada Pérez, Ana & Diego Pascual y Cabo. 2012. Interface heritage speech across proficiencies: Unaccusativity, focus, and subject position in Spanish. In Kimberly Geeslin & Manuel Díaz-Campos (eds.), *Selected proceedings from the Hispanic linguistics symposium 2010*, 308–318. Somerville, MA: Cascadilla Proceedings Project. Cited on page 182.
- de Vogelaer, Gunther & Matthias Katerbow. 2017. *Acquiring sociolinguistic variation*. Amsterdam: John Benjamins Publishing Company. Cited on page 78.
- Deuchar, Margaret. 2006. Welsh-English code-switching and the Matrix Language Frame model. *Lingua* 116(11), 1986–2011. doi:10.1016/j.lingua.2004.10.001. Cited on pages 5 and 72.
- Deuchar, Margaret, Peredur Davies & Kevin Donnelly. 2018. *Building and using the Siarad corpus: Bilingual conversations in Welsh and English*. Amsterdam: John Benjamins Publishing Company. Cited on pages 5, 28, 30, 35, 36, 38, 40, 72, and 96.
- Diab, Mona & Ankit Kamboj. 2011. Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for Hindi English code switched data: A pilot annotation. In *Proceedings of the ninth workshop on Asian language resources*, 36–40. Chiang Mai, Thailand: Asian Federation of Natural Language Processing. <https://www.aclweb.org/anthology/W11-3407>. Cited on page 48.
- Diep, Quang Ban. 2004. *Ngữ pháp tiếng việt [Vietnamese grammar]*. Ha Noi: Nhà Xuất Bản Giáo Dục. Cited on pages 130 and 183.
- Dijkstra, Ton & Walter van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition* 5(3), 175–197. doi:10.1017/S1366728902003012. Cited on page 184.
- Do, Hoa, Huu Thuy Giang Tran & Ket Mai. 2018. Vietnamese telephone openings: Both universals and particulars. *Language and Dialogue* 8(3), 363–389. doi:10.1075/ld.00022.do. Cited on page 126.
- Dorian, Nancy. 2014. Language loss and maintenance in language contact situations. In Barbara F Freed & Richard D. Lambert (eds.), *The loss of language skills*, 203–222. Leiden, The Netherlands: Brill. doi:10.1163/9789004261938\_013. Cited on page 5.
- Dorr, Bonnie J., Matt Snover & Nitin Madnani. 2011. Machine translation evaluation and optimization. In Joseph Olive, Caitlin Christianson & John McCary (eds.), *Handbook of natural language processing and machine translation: DARPA Global autonomous language exploitation*, New York: Springer. Cited on page 57.

- Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. Outline of discourse transcription. In Jane Edwards & Martin D. Lampert (eds.), *Talking data: Transcription and coding in discourse research*, Hillsdale, NJ: Lawrence Erlbaum. Cited on pages 39, 41, and 233.
- Dubinina, Irina & Maria Polinsky. 2013. Russian in the USA. In Michael Moser & Maria Polinsky (eds.), *Slavic Languages in Migration (Slavische Sprachgeschichte)*, 131–160. Wien: Lit Verlag. Cited on page 177.
- Duffield, Nigel. 2009. Head-first: On the head-initiality of Vietnamese clauses. In *Workshop on linguistics of Vietnamese*, University of Stuttgart. Cited on page 88.
- Eckert, Penelope. 2004. The meaning of style. In Wai-Fong Chiang, Elaine Chun, Laura Mahalingappa & Siri Mehus (eds.), *Symposium about Language and Society, SALSA 11*, 41–53. Austin, TX: Texas Linguistics Forum 47. Cited on pages 122, 164, and 205.
- Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4), 453–476. doi:10.1111/j.1467-9841.2008.00374.x. Cited on page 122.
- Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41, 87–100. doi:10.1146/annurev-anthro-092611-145828. Cited on pages 120 and 122.
- Eckert, Penelope & Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* doi:10.1146/annurev.an.21.100192.002333. Cited on pages 6, 20, 21, and 122.
- Emanau, Murray B. 1951. *Studies in Vietnamese (Annamese) grammar*. Berkeley / Los Angeles: University of California Press. Cited on page 48.
- Erickson, Jon. 2001. English. In Jane Garry, Carl R. Galvez Rubino, H. W. Wilson, Adams B. Bodomo, Robert French & Alice Faber (eds.), *Facts about the world's languages: An encyclopedia of the world's major languages, past, and present*, 199–203. New York: H. W. Wilson Company. Cited on page 89.
- Finlayson, Rosalie, Karen Calteaux & Carol Myers-Scotton. 1998. Orderly mixing and accommodation in South African codeswitching. *Journal of Sociolinguistics* doi:10.1111/1467-9481.00052. Cited on page 72.
- Flores-Ferrán, Nydia. 2002. *Subject personal pronouns in Spanish narratives of Puerto Ricans in New York City: A sociolinguistic perspective*. Munich: Lincom Europa. Cited on page 147.

- Ford, Cecilia E. & Sandra A. Thompson. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In Elinor Ochs, Emanuel A. Schegloff & Sandra Thompson (eds.), *Interaction and Grammar*. Studies in Interactional Sociolinguistics, 134–184. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620874.003. Cited on page 42.
- Frascarelli, Mara. 2018. The interpretation of pro in consistent and partial null-subject languages: A comparative interface analysis. In Federica Cognola & Jan Casalicchio (eds.), *Null Subjects in Generative Grammar: A Synchronic and Diachronic Perspective*, 211–239. Oxford: Oxford University Press. Cited on pages 150 and 207.
- Frey, J. William. 1942. *A simple grammar of Pennsylvania Dutch*. Lancaster, PA: Brookshire Publications. Cited on page 80.
- Fried, Mirjam & Jan-Ola Östman. 2005. Construction grammar and spoken language: The case of pragmatic particles. *Journal of Pragmatics* 37(11), 1752–1778. Cited on page 96.
- Fuertes, Raquel Fernández, Juana Licerias & Anahí Alba de la Fuente. 2013. Beyond the subject DP versus the subject pronoun divide in agreement switches. In Christina Tortora, Marcel den Dikken, Ignacio L. Montoya & Teresa O'Neill (eds.), *Romance Linguistics 2013: Selected papers from the 43rd linguistic symposium on Romance languages (LSRL)*, 79–98. Amsterdam: John Benjamins Publishing Company. Cited on page 99.
- Fuller, Janet M. 1996. When cultural maintenance means linguistic convergence: Pennsylvania German evidence for the Matrix Language Turnover Hypothesis. *Language in Society* 25(4), 493–514. Cited on pages 79, 80, 83, and 101.
- Fuller, Janet M. & Heike Lehnert. 2000. Noun phrase structure in German-English codeswitching: Variation in gender assignment and article use. *International Journal of Bilingualism* 4(3), 399–420. doi:10.1177/13670069000040030601. Cited on pages 5 and 72.
- Gal, Susan. 1978. Peasant men can't get wives: Language change and sex roles in a bilingual community. *Language in Society* 7(1), 1–16. doi:10.1017/S0047404500005303. Cited on page 28.
- Gal, Susan. 1979. *Language shift: Social determinants of linguistic change in bilingual Austria*. New York: Academic Press. Cited on page 28.
- Garcia-Colon, Ismael. 2004. Legacies: The story of the immigrant second generation. *American Anthropologist* 106(2), 391–395. doi:10.1525/aa.2004.106.2.391.1. Cited on page 107.

- Gardner-Chloros, Penelope. 1991. *Language selection and switching in Strasbourg*. Oxford: Oxford University Press. Cited on page 51.
- Gardner-Chloros, Penelope. 2009. *Code-switching*. Cambridge: Cambridge University Press. Cited on pages 51, 65, 71, and 83.
- Gardner-Chloros, Penelope & Malcolm Edwards. 2004. Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring. *Transactions of the Philological Society* 102(1), 103–129. doi:10.1111/j.0079-1636.2004.00131.x. Cited on pages 83 and 99.
- Garrett, Peter. 2010. *Attitudes to language*, vol. 9780521766. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511844713. Cited on page 30.
- Geeslin, Kimberly & Pedro Guijarro-Fuentes. 2008. Variation in contemporary Spanish: Linguistic predictors of *estar* in four cases of language contact. *Bilingualism: Language and Cognition* 11(3), 365–380. Cited on page 154.
- Ginzburg, Jonathan & Massimo Poesio. 2016. Grammar is a system that characterizes talk in interaction. *Frontiers in psychology* 7(1938). doi:10.3389/fpsyg.2016.01938. Cited on page 96.
- Givón, Talmy. 1983. Topic continuity in discourse: An introduction. In Kimberly Geeslin & Manuel Diaz-Campos (eds.), *Topic continuity in discourse: A quantitative cross-linguistic study*, 1–41. Amsterdam: John Benjamins Publishing Company. Cited on page 149.
- Goldberg, Adele. 2001. Patient arguments of transitive verbs can be omitted: The role of information structure in argument distribution. *Language Sciences* 23, 503–524. Cited on page 128.
- Greco, Ciro, Trang Phan & Liliane Haegeman. 2018. On *nó* as an optional expletive in Vietnamese. In Federica Cognola & Jan Casalicchio (eds.), *Null subjects in generative grammar: A synchronic and diachronic perspective*, 31–51. Oxford: Oxford University Press. doi:10.1093/oso/9780198815853.003.0002. Cited on page 140.
- Green, David. 1998. Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition* 1(2), 67–82. Cited on page 184.
- Grégoire, Antoine. 1971. L'apprentissage du langage. In Aaron Bar-Adon & Werner F. Leopold (eds.), *Child language: A book of readings*, 91–95. Englewood Cliffs, NJ: Prentice-Hall. Cited on page 113.

- Gumperz, John. 1968. The speech community. In David L. Sills & Robert K. Merton (eds.), *International encyclopedia of social sciences*, 381–386. New York: Macmillan Reference. Cited on page 18.
- Gumperz, John. 1977. The sociolinguistic significance of conversational code-switching. *RELC Journal* 8(2), 1–34. doi:10.1177/003368827700800201. Cited on page 99.
- Gutierrez, Manuel. 2003. Simplification and innovation in US Spanish. *Multilingua* 22(2), 169–184. Cited on page 179.
- Ha, Kieu-Phuong. 2012. *Prosody in Vietnamese: Intonational form and function of short utterances in conversation*. Canberra: Asia-Pacific Linguistics (SEAMLES). Cited on page 129.
- Habtoor, Hussein Ali. 2012. Language maintenance and language shift among second generation tigrinya-speaking eritrean immigrants in Saudi Arabia. *Theory and Practice in Language Studies* 2(5), 945–955. doi:10.4304/tpls.2.5.945-955. Cited on page 69.
- Haeri, Niloofar. 1998. Overt and non-overt subjects in Persian. *IPrA Papers in Pragmatics* 3(1), 155–166. doi:10.1075/iprapip.3.1.05hae. Cited on page 150.
- Halliday, Michael. 1967. *Intonation and grammar in British English*. Janua Linguarum. Series Practica; 48. The Hague: Mouton. Cited on page 41.
- Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers Linguistiques d' Ottawa* 16, 15–25. Cited on page 148.
- Haugen, Einar. 1950. The analysis of linguistic borrowing. *Language* 26(2), 210–231. doi:10.2307/410058. Cited on page 111.
- Hawkins, Roger. 2000. Persistent selective fossilisation in second language acquisition and the optimal design of the language faculty. *Essex Research Reports in Linguistics* 34, 75–90. Cited on page 116.
- Hay, Jennifer. 2011. Statistical analysis. In Marianna Di Paolo & Malcah Yaeger-Dror (eds.), *Sociophonetics: A student's guide*, 198–214. Abingdon: Routledge. Cited on page 158.
- Hernández, José Esteban. 2009. Measuring rates of word-final nasal velarization: The effect of dialect contact on in-group and out-group exchanges. *Journal of Sociolinguistics* 13(5), 583–612. Cited on page 146.

- Herring, Jon Russell, Margaret Deuchar, M. Carmen Parafita Couto & Mónica Quintanilla Moro. 2010. 'I saw the madre': Evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data. *Journal of Bilingual Education and Bilingualism* 13(5), 553-573. Cited on page 75.
- Hinds, John. 1975. Third person pronouns in Japanese. In Fred C. Peng (ed.), *Language in Japanese Society*, 129-175. Tokyo: University of Tokyo Press. Cited on page 126.
- Hinds, John. 1983. Topic continuity in Japanese. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study*, 47-93. Amsterdam: John Benjamins Publishing Company. Cited on page 126.
- Hinskens, Frans & Peter Auer. 1997. The role of interpersonal accommodation in a theory of language change. In Peter Auer, Frans Hinskens & Paul Kerswill (eds.), *Dialect change: The convergence and divergence of dialects in contemporary societies*, 335-357. Cambridge: Cambridge University Press. Cited on page 210.
- Hoffman, Michol F. & James A. Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change* 22(1), 37-67. doi: 10.1017/S0954394509990238. Cited on page 23.
- Holm, John. 1984. Variability of the copula in Black English and its creole kin. *American Speech* 59(4), 291-309. doi:10.2307/454782. Cited on page 134.
- Hu, Qian. 1993. The acquisition of classifiers by young Mandarin-speaking children. Boston University Phd dissertation. Cited on page 111.
- Huang, C-T James. 1984. On the distribution and reference of empty pronouns. *Linguistic Inquiry* 15(4), 531-74. Cited on pages 128, 151, 152, and 170.
- Hudson Kam, Carla. 2015. The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language* 91(4), 906-37. Cited on page 78.
- Huffman, Franklin & Trong Hai Tran. 2004. *Intermediate spoken Vietnamese*. Ithaca: Cornell University South East Asian Programme. Cited on page 129.
- Hurewitz, Felicia. 1998. A quantitative look at discourse coherence: In centering theory in discourse. In Marilyn A Walker, Arvind K Joshi & Ellen F Prince (eds.), *Centering theory in discourse*, 273-291. Oxford: Clarendon Press. Cited on page 44.

- Ihemere, Kelechukwu. 2016. In support of the Matrix Language Frame model: Evidence from Igbo-English intrasentential code-switching. *Language Matters* 47(1), 105–127. Cited on page 72.
- Ihemere, Kelechukwu. 2017. Igbo-English intrasentential codeswitching and the Matrix Language Frame model. In Doris L. Payne, Sara Pacchiarotti & Mokaya Bosire (eds.), *Diversity in African languages*, 539–559. Berlin: Language Science Press. doi:10.17169/langsci.b121.498. Cited on page 72.
- Ishizawa, Hiromi. 2004. Minority language use among grandchildren in multigenerational households. *Sociological Perspectives* 47(4). doi:10.1525/sop.2004.47.4.465. Cited on page 107.
- Ivanova-Sullivan, Tania. 2014. *Theoretical and experimental aspects of syntax-discourse interface in heritage grammars*. Leiden, The Netherlands: Brill. <https://brill.com/view/title/23544>. Cited on page 176.
- Jake, Janice L. & Carol Myers-Scotton. 2009. Second generation shifts in sociopragmatic orientation and code-switching patterns. In Aleya Rouchdy (ed.), *Language contact and language conflict in Arabic*, 317–330. New York: Routledge. Cited on pages 96 and 115.
- Jia, Li & Robert Bayley. 2002. Null pronoun variation in Mandarin Chinese. *University of Pennsylvania Working Papers in Linguistics* 8(3), 103–116. Cited on pages 147, 149, 187, 188, and 207.
- 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. doi:10.1111/j.1749-818X.2008.00108.x. Cited on pages 157, 158, and 159.
- Johnstone, Barbara & Scott F. Kiesling. 2008. Indexicality and experience: Exploring the meanings of /aw/-monophthongization in Pittsburgh. *Journal of Sociolinguistics* 12(1), 5–33. doi:10.1111/j.1467-9841.2008.00351.x. Cited on page 155.
- Joshi, Aravind. 1985. Processing sentences with intrasentential code-switching. In D. R. Dowty, L. Karttunen & A. Zwicky (eds.), *Natural Language Parsing*, 190–205. Cambridge: Cambridge University Press. Cited on pages 4 and 64.
- Jung, Dagmar & Nikolaus Himmelmann. 2011. Retelling data: Working on transcription. In Geoffrey Haig, Nicole Nau, Stefan Snell & Claudia Wegener (eds.), *Documenting endangered languages: Achievements and perspectives*, 201–222. Berlin: de Gruyter. Cited on page 39.

- Kaltsa, Maria, Ianthi Tsimpli & Jason Rothman. 2015. Exploring the source of differences and similarities in L1 attrition and heritage speaker competence: Evidence from pronominal resolution. *Lingua* 164(B), 266–288. doi:10.1016/j.lingua.2015.06.002. Cited on page 197.
- Kapferer, Bruce & Barry Morris. 2003. The Australian society of the state: Egalitarian ideologies and new directions in exclusionary practice. *Social Analysis* 47(3), 80–107. doi:10.3167/015597703782352835. Cited on page 164.
- Karidakis, Maria & Dharma Arunachalam. 2016. Shift in the use of migrant community languages in Australia. *Journal of Multilingual and Multicultural Development* 37(1). doi:10.1080/01434632.2015.1023808. Cited on page 107.
- Kautzsch, Alexander. 2012. *The historical evolution of earlier African American English*. Berlin: de Gruyter. doi:10.1515/9783110907971. Cited on page 155.
- Kay, Paul & Chad K. McDaniel. 1979. On the logic of variable rules. *Language in Society* 8(2), 151–187. doi:10.1017/S0047404500007429. Cited on page 158.
- Kheir, Afifa Eve. 2019. The Matrix Language Turnover Hypothesis: The case of the Druze language in Israel. *Journal of Language Contact* 12(2), 479–512. doi:10.1163/19552629-01202008. Cited on pages 80, 81, and 82.
- Kidwai, Sana, Christopher Bryant, Li Nguyen & Theresa Biberauer. 2019. Automatic language identification in code-switched Hindi-English social media texts. In *Cambridge language sciences symposium: Perspectives on language change*, Cambridge. Cited on page 59.
- Kiesling, Scott. 2005. Variation, stance and style: Word-final -er, high rising tone, and ethnicity in Australian English. *English World-Wide* 26(1), 1–42. doi:10.1075/eww.26.1.02kie. Cited on pages 19, 20, 23, and 120.
- Kiesling, Scott. 2009. Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In Marianna Di Paolo & Malcah Yaeger-Dror (eds.), *Stance: Sociolinguistic Perspectives*, 171–194. Oxford: Oxford University Press. Cited on page 155.
- Kiesling, Scott. 2011. *Linguistic variation and change*. Edinburgh: Edinburgh University Press. Cited on pages 18, 19, 120, 124, 155, and 168.
- Kim, Haeyeon. 1989. Nominal reference in discourse: Introducing and tracking referents in Korean spoken narratives. *Harvard Studies in Korean Linguistics* 3, 431–444. Cited on page 128.

- King, Kendall & Lyn Fogle. 2006. Bilingual parenting as good parenting: Parents' perspectives on family language policy for additive bilingualism. *International Journal of Bilingual Education and Bilingualism* 9(6), 695–712. doi:10.2167/beb362.0. Cited on pages 16 and 107.
- Kipp, Sandra, Michael Clyne & Anne Pauwels. 1999. *Immigration and Australia's language resource*. Canberra: Australian Government Publishing Service. Cited on page 15.
- Koehn, Philipp. 2009. *Statistical machine translation*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815829. Cited on page 57.
- Koornneef, Arnout, Sergey Avrutin, Frank Wijnen & Eric Reuland. 2011. Tracking the preference for bound variable dependencies in ambiguous ellipses and Only-structures. In Jeffrey Runner (ed.), *Experiments at the interfaces*, 67–100. Bingley: Emerald. Cited on page 182.
- Labov, William. 1966. *The social stratification of English in New York City*. Cambridge: Cambridge University Press. Cited on pages 122 and 155.
- Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45(4), 715–762. doi:10.2307/412333. Cited on pages 80, 154, and 155.
- Labov, William. 1972. Some principles of linguistic methodology. *Language in Society* 1(1), 97–120. doi:10.1017/S0047404500006576. Cited on pages vi, 4, 5, 6, 18, 22, 80, 119, 120, 134, 137, 155, 203, and 204.
- Labov, William. 1982. Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. *Language in Society* 11(2), 165–201. Cited on page 45.
- Labov, William. 1984. Field methods of the project on linguistic change and variation. In John Baugh & Joel Sherzer (eds.), *Language in use: Readings in sociolinguistics*, 28–53. Englewood Cliffs, NJ: Prentice Hall. Cited on pages 33, 123, and 155.
- Labov, William. 1995. The case of the missing copula: The interpretation of zeros in African American English. In Lila Gleitman & Mark Liberman (eds.), *Language: An invitation to Cognitive Science*, 25–54. Cambridge, MA: MIT Press. Cited on page 78.
- Labov, William, Paul Cohen, Clarence Robins & John Lewis. 1968. *A study of the non-standard English of Negro and Puerto Rican speakers in New York City*. Philadelphia: U.S. Regional Survey: Co-operative Research Report 3288, Volume 1. Cited on page 154.
- Laleko, Oksana & Maria Polinsky. 2016. Between syntax and discourse: Topic and case marking in heritage speakers and L2 learners of Japanese and Korean. *Linguistic Approaches to Bilingualism* 6(4), 396–439. doi:10.1075/lab.14018.lal. Cited on page 177.

- Laleko, Oksana & Maria Polinsky. 2017. Silence is difficult: On missing elements in bilingual grammars. *Zeitschrift für Sprachwissenschaft* 36(1), 135–163. doi:10.1515/zfs-2017-0007. Cited on pages 177 and 179.
- Landa, Alazne. 1995. Conditions on null objects in Basque Spanish and their relation to “leísmo” and clitic doubling. University of Southern California Phd dissertation. Cited on page 153.
- Lardiere, Donna. 1998. Dissociating syntax from morphology in a divergent L2 end-state grammar. *Second Language Research* 14(4), 359–375. doi:10.1191/026765898672500216. Cited on page 85.
- Lardiere, Donna. 2007. *Ultimate attainment in second language acquisition*. New York: Routledge. Cited on page 116.
- Lavie, Alon & Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, doi:10.1007/s10590-009-9059-4. Cited on page 57.
- Le, Phuc Thien. 2011. Transnational variation in linguistic politeness in Vietnamese: Australia and Vietnam. Melbourne: Victoria University Phd dissertation. Cited on page 126.
- Le, Thao. 1995. Literacy challenges and the Vietnamese communities in Australia. In David Myers (ed.), *Reinventing literacy: the multicultural imperative*, Brisbane: Watson Ferguson. Cited on page 15.
- Le Page, Robert Brock. 1989. What is a language? *York Papers in Linguistics* 13, 9–24. Cited on page 121.
- Le Page, Robert Brock & Andrée Tabouret-Keller. 1985. *Acts of identity: Creole-based approaches to language and ethnicity*. Cambridge: Cambridge University Press. Cited on pages 121 and 122.
- Lecanda, Lorena Sainz-Maza & Scott Schwenter. 2017. Null objects with and without bilingualism in the Portuguese- and Spanish-speaking world. In Kate Bellamy, Michael Child, Paz González, Antje Muntendam & Maria del Carmen Parafita Couto (eds.), *Multidisciplinary approaches to bilingualism in the Hispanic and Lusophone world*, 95–119. Amsterdam: John Benjamins Publishing Company. doi:10.1075/ihll.13.05sai. Cited on page 153.
- Lee, Boh Young. 2012. Heritage language maintenance and cultural identity formation: The case of Korean immigrant parents and their children in the USA. *Early Child Development and Care* 183(11), 1576–1588. doi:10.1080/03004430.2012.741125. Cited on page 16.

- Lee, Duck-Young & Yoko Yonezawa. 2008. The role of the overt expression of first and second person subject in Japanese. *Journal of Pragmatics* 40(4), 733–767. doi:10.1016/j.pragma.2007.06.004. Cited on page 149.
- Leech, Geoffrey. 2000. Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4), 675–724. doi:10.1111/0023-8333.00143. Cited on page 96.
- Leech, Geoffrey & Jan Svartvik. 2003. *A communicative grammar of English*. Abingdon: Routledge. Cited on page 96.
- Li, Charles N. & Sandra A. Thompson. 1979. Third person pronouns and zero anaphora in Chinese discourse. In Talmy Givón (ed.), *Syntax and semantics: Discourse and syntax*, 311–335. New York: Academic Press. Cited on page 149.
- Li, Xiaoshi, Xiaoqing Chen & Wen-Hsin Chen. 2012. Variation of subject pronominal expression in Mandarin Chinese. *Sociolinguistic Studies* 6(1), 91–119. doi:10.1558/sols.v6i1.91. Cited on page 149.
- Li, Xiaoting. 2014. *Multimodality, interaction and turn-taking in Mandarin conversation*. Amsterdam: John Benjamins Publishing Company. Cited on page 41.
- 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. doi:10.1191/026765899678128123. Cited on page 149.
- Liddicoat, Anthony. 2018. Indigenous and immigrant languages in Australia. In Corinne A. Seals & Sheena Shah (eds.), *Heritage language policies around the world*, 237–253. New York: Routledge. Cited on page 2.
- Lippi-Green, Rosina L. 1989. Social network integration and language change in progress in a rural alpine village. *Language in Society* 18(2), 213–234. doi:10.1017/S0047404500013476. Cited on page 20.
- Lipski, John. 1978. Code-switching and the problem of bilingual competence. In Michel Paradis (ed.), *Aspects of bilingualism*, 250–264. Columbia, SC: Hornbeam Press. Cited on page 99.
- Lipski, John. 2019. Field-testing code-switching constraints: A report on a strategic languages project. *Languages* 4(1), 1–29. doi:10.3390/languages4010007. Cited on page 99.
- Lo, Chi-kiu & Dekai Wu. 2013. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In *Proceedings of the Eighth Workshop on Statis-*

- tical Machine Translation*, 422–428. Association for Computational Linguistics. Cited on page 57.
- Long, Michael H. 2008. Stabilization and fossilization in interlanguage development. In Michael H. Long & Catherine J. Doughty (eds.), *The handbook of second language acquisition*, 487–535. Hoboken, NJ: John Wiley & Sons, Ltd. doi:10.1002/9780470756492.ch16. Cited on page 116.
- Lozano, Cristobal. 2006. Focus and split-intransitivity: The acquisition of word order alternations in non-native Spanish. *Second Language Research* 22(2), 145–187. doi:10.1191/0267658306sr264oa. Cited on pages 132 and 177.
- Lustres, Eduardo. 2018. The acquisition of obligatory and variable subjunctive mood selection in temporal and concessive clauses in heritage and L2 Spanish. Purdue University, Indiana Phd dissertation. Cited on page 177.
- Lyu, Dau-Cheng, Tien-Ping Tan, Eng-Siong Chng & Haizhou Li. 2015. Mandarin-English code-switching speech corpus in South-East Asia: SEAME. In *Language Resources and Evaluation*, vol. 49, 581–600. Cited on page 48.
- MacSwan, Jeff. 2005. Codeswitching and generative grammar: A critique of the MLF model and some remarks on “modified minimalism”. *Bilingualism: Language and Cognition* 8(1), 1–22. doi:10.1017/S1366728904002068. Cited on pages 74, 75, and 99.
- MacSwan, Jeff & Sonia Colina. 2014. Some consequences of language design: Codeswitching and the PF interface. In Jeff MacSwan (ed.), *Grammatical theory and bilingual codeswitching*, 185–210. Cambridge, MA: MIT Press. Cited on page 99.
- Mair, Christian. 2013. Writing the corpus-based history of spoken English: The elusive past of a cleft construction. In Gisle Andersen & Kristin Bech (eds.), *English Corpus Linguistics: Variation in Time, Space and Genre. Selected papers from ICAME 32*, vol. 77, 11–29. Leiden, The Netherlands: Brill. doi:10.1163/9789401209403\_003. Cited on page 96.
- Malik, Nazir Ahmed & Muhammad Ajmal Khurshid. 2017. Empirical inadequacy of the functional head constraint: Evidence from Urdu/English code-switching. *Kashmir Journal of Language Research* 20(2), 87–101. Cited on page 99.
- Marcus, Gary, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu & Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* 57(4), i–178. doi:10.2307/1166115. Cited on page 114.

- Margaza, Panagiota & Aurora Bel. 2006. Null subjects at the syntax-pragmatics interface: Evidence from Spanish interlanguage of Greek speakers. In Mary Grantham O'Brien, Christine Shea & John Archibald (eds.), *Proceedings of the 8th generative approaches to second language acquisition conference*, 88–97. Somerville, MA: Cascadilla Press. Cited on pages 132 and 177.
- Marian, Viorica & Michael Spivey. 2003. Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition* 6(2), 97–115. doi:10.1017/S1366728903001068. Cited on page 184.
- Mayeux, Oliver. 2019. Rethinking decreolization: Language contact and change in Louisiana creole. University of Cambridge Phd dissertation. Cited on page 114.
- McConvell, Patrick. 2010. Contact and indigenous languages in Australia. In Raymond Hickey (ed.), *The handbook of language contact*, 770–794. New York: Wiley-Blackwell. Cited on page 134.
- Mesthrie, Rajend & Rakesh M. Bhatt. 2008. *World Englishes: The study of new linguistic varieties. Key topics in sociolinguistics*. Cambridge: Cambridge University Press. Cited on page 78.
- 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. doi:10.1017/S0954394509990196. Cited on pages 150 and 153.
- Michaud, Alexis & Marc Brunelle. 2014. Information structure in Asia Yongning Na (Sino-Tibetan) and Vietnamese (austroAsiatic). In Caroline Féry & Shinichiro Ishihara (eds.), *The Oxford handbook of information structure*, Oxford: Oxford University Press. Cited on page 126.
- Milroy, Lesley. 1980. *Language and social networks*. Oxford: Basil Blackwell. Cited on pages 17, 20, and 27.
- Milroy, Lesley & James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society* 21(1), 1–26. Cited on page 155.
- Milroy, Lesley & Li Wei. 1995. A social network approach to code-switching: The example of a bilingual community in Britain. In Lesley Milroy & Pieter Muysken (eds.), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, 136–157. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620867.007. Cited on pages 27, 28, and 120.
- Mobaraki, Mohsen, Anne Vainikka & Martha Young-Scholten. 2008. The status of subjects in early child L2 English. In Belma Haznedar & Elena Gavruseva (eds.), *Current trends in child*

- second language acquisition: A generative perspective*, 209–235. Amsterdam: John Benjamins Publishing. doi:10.1075/lald.46.11mob. Cited on page 155.
- Mohring, Anja & Jurgen Meisel. 2003. The verb-object parameter in simultaneous and successive acquisition of bilingualism. In Jochen Rehbein, Christiane Hohenstein & Lukas Pietsch (eds.), *(In)vulnerable domains in multilingualism*, 295–334. Amsterdam: John Benjamins Publishing Company. doi:10.1075/hsm.1.11moh. Cited on page 177.
- Montrul, Silvina. 2002. Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and Cognition* 5(1), 39–68. doi:10.1017/S1366728902000135. Cited on page 176.
- Montrul, Silvina. 2004. Subject and object expression in Spanish heritage speakers: A case of morpho-syntactic convergence. *Bilingualism: Language and Cognition* 7(2), 125–142. doi:10.1017/S1366728904001464. Cited on pages 176 and 177.
- Montrul, Silvina. 2005. Second language acquisition and first language loss in adult early bilinguals: Exploring some differences and similarities. *Second Language Research* 21(3), 199–249. doi:10.1191/0267658305sr247oa. Cited on page 177.
- Montrul, Silvina. 2006. On the bilingual competence of Spanish heritage speakers: Syntax, lexical-semantics and processing. *International Journal of Bilingualism* 10(1), 37–69. doi:10.1177/13670069060100010301. Cited on page 177.
- Montrul, Silvina. 2008. *Incomplete acquisition in bilingualism. Re-examining the age factor*. Amsterdam: John Benjamins Publishing Company. doi:10.1017/9781107252349. Cited on page 176.
- Montrul, Silvina. 2015. *The acquisition of heritage languages*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139030502. Cited on pages 2, 176, and 206.
- Montrul, Silvina & Noelia Sánchez-Walker. 2013. Differential object marking in child and adult Spanish heritage speakers. *Language Acquisition* 20(2), 109–132. doi:10.1080/10489223.2013.766741. Cited on page 133.
- Mooney, Damien. 2018. Quantitative approaches for modelling variation and change: A case study of sociophonetic data from Occitan. In Wendy Ayres-Bennett & Janice Carruthers (eds.), *Manuals of Romance linguistics*, vol. 18, 59–90. Berlin: de Gruyter. Cited on pages 158 and 159.

- Moorkens, Joss, Sheila Castilho, Federico Gaspari & Stephen Doherty (eds.). 2018. *Translation quality assessment: From principles to practice*. New York: Springer. doi:10.1007/978-3-319-91241-7. Cited on page 57.
- Müller, Natascha. 2007. Some notes on the syntax-pragmatics interface in bilingual children: German in contact with French/Italian. In Jochen Rehbein, Christiane Hohenstein & Lukas Pietsch (eds.), *Connectivity in grammar and discourse*, 101–135. Amsterdam: John Benjamins Publishing Company. doi:10.1075/hsm.5.07mu. Cited on page 177.
- Müller, Natascha & Aafke Hulk. 2001. Cross-linguistic influence in bilingual language acquisition: Italian and French as recipient languages. *Bilingualism: Language and Cognition* 4(1), 1–21. doi:10.1017/S1366728901000116. Cited on page 176.
- Muysken, Pieter. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge: Cambridge University Press. Cited on pages 104 and 134.
- Myers-Scotton, Carol. 1993. *Duelling languages: Grammatical structure in codeswitching*. Oxford: Clarendon. Cited on pages v, 4, 5, 51, 64, 65, 67, 72, 73, 76, 79, 99, 102, 103, and 203.
- Myers-Scotton, Carol. 1997. *Duelling languages: Grammatical structure in codeswitching (revised with a new afterword)*. Oxford: Clarendon Press. Cited on page 65.
- Myers-Scotton, Carol. 1998. A way to dusty death: The Matrix Language Turnover Hypothesis. In Lenore A. Grenoble & Lindsay J. Whaley (eds.), *Endangered languages: Language loss and community response*, Cambridge: Cambridge University Press. doi:10.1017/CBO9781139166959.013. Cited on pages v, 63, 68, 69, 70, 71, 79, 96, 106, 108, and 204.
- Myers-Scotton, Carol. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press. Cited on pages v, 4, 64, 65, 66, 67, 68, 69, 73, 75, 76, 78, 79, 81, 84, 99, 101, 102, 104, 107, 110, 111, and 115.
- Myers-Scotton, Carol. 2003. What lies beneath: Split (mixed) languages as contact phenomena. In Yaron Matras & Peter Bakker (eds.), *Trends in linguistics. The mixed language debate: Theoretical and empirical Advances*, 73–106. Berlin: de Gruyter. Cited on page 71.
- Myers-Scotton, Carol. 2004. Research note and erratum. *Bilingualism: Language and Cognition* 7(1), 89–90. doi:10.1017/S1366728904001294. Cited on page 73.
- Myers-Scotton, Carol. 2006. *Multiple voices: An introduction to bilingualism*. Malden & Oxford: Blackwell Publishing Ltd. Cited on pages 5, 72, and 92.

- Myers-Scotton, Carol & Janice L. Jake. 1995. Matching lemmas in a bilingual language competence and production model: evidence from intrasentential code-switching. *Linguistics* 33(5), 981–1024. doi:10.1515/ling.1995.33.5.981. Cited on page 70.
- Myers-Scotton, Carol & Janice L. Jake. 2000a. Four types of morpheme: Evidence from aphasia, code-switching, and second-language acquisition. *Linguistics* 38(6). doi:10.1515/ling.2000.021. Cited on pages 66 and 77.
- Myers-Scotton, Carol & Janice L. Jake. 2000b. Testing the 4-M model: An introduction. *International Journal of Bilingualism* 4(1). doi:10.1177/13670069000040010101. Cited on page 66.
- Myers-Scotton, Carol & Janice L. Jake. 2009. A universal model of code-switching and bilingual language processing and production. In Barbara Bullock & Almeida Jacqueline Toribio (eds.), *Cambridge handbooks in linguistics. The Cambridge handbook of linguistic code-switching*, 336–357. Cambridge: Cambridge University Press. Cited on page 68.
- Nagy, Naomi. 2011. A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana di Linguistica Applicata* 43(1-2), 65–84. Cited on pages 34 and 38.
- Nagy, Naomi. 2015. A sociolinguistic view of null subjects and vot in Toronto heritage languages. *Lingua* 164, 309–327. doi:10.1016/j.lingua.2014.04.012. <http://www.sciencedirect.com/science/article/pii/S0024384114001004>. Fundamentally (in)complete grammars? Emergence, acquisition and diffusion of new varieties. Cited on pages 132, 133, 146, 168, and 169.
- Nagy, Naomi, 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. Cited on pages 147, 149, and 150.
- Nagy, Naomi, Michael Iannozzi & David Heap. 2017. Faetar null subjects: A variationist study of a heritage language in contact. *International Journal of the Sociology of Language* 2018(249), 31–47. doi:10.1515/ijsl-2017-0040. Cited on page 177.
- Nagy, Naomi & Devyani Sharma. 2013. Transcription. In Robert Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 235–256. Cambridge: University of Cambridge. Cited on pages 38 and 39.
- Nakamura, Masaru. 1991. Japanese as a pro language. *The Linguistic Review* 6(4). doi:10.1515/tlir.1987.6.4.281. Cited on page 128.
- Namba, Kazuhiko. 2004. An overview of Myers-Scotton's Matrix Language Frame model. *Senri International School (SIS) Educational Research Bulletin* 9, 1–10. Cited on page 115.

- National Australia Bank. 2017. NAB charitable giving index: Insights into the donating behaviours of Australian consumers. Tech. rep. NAB Canberra, ACT. Cited on page 17.
- National Indochinese Clearinghouse, Center for Linguistics. 1977. *English pronunciation exercises for speakers of Vietnamese*. Arlington, VA: Center for Applied Linguistics. Cited on page 85.
- Ngo, Binh. 2019. Vietnamese pronouns in discourse. University of Southern California Phd dissertation. Cited on pages 150, 187, 188, and 207.
- Ngo, Binh. 2020. *Vietnamese: An essential grammar*. New York: Routledge. Cited on page 125.
- Ngo, Thanh. 2006. Translation of Vietnamese terms of address and reference. *Translation Journal* 10(4), Online publishing. <https://translationjournal.net/journal/38viet.htm>. Cited on pages 125 and 126.
- Nguyen, Dinh Hoa. 1957. Classifiers in Vietnamese. *Word* 13(1), 124–52. doi:10.1080/00437956.1957.11659631. Cited on pages 109, 111, and 129.
- Nguyen, Dinh Hoa. 1997. *Vietnamese*. Amsterdam: John Benjamins Publishing Company. Cited on pages 48, 90, 94, 101, 110, 124, 125, 126, 127, 128, 129, and 142.
- Nguyen, Kim Than. 1975. An outline of Vietnamese grammar. *Vietnamese Studies* 40, 148–217. Cited on pages 129 and 131.
- Nguyen, Li. 2015. 'It just makes more sense if I try than forcing my parents to learn another language' - a pilot study on Vietnamese migrants' language attitudes across two generations. Unpublished postgraduate study. The Australian National University. Cited on pages 15, 16, and 155.
- Nguyen, Li. 2016. Incorporated kin terms, bilingual speakers: Probing Vietnamese-English bilingual speech via discourse distribution and pragmatic norms. The Australian National University masters thesis. Cited on pages 38, 51, 81, and 103.
- Nguyen, Li. 2018. Borrowing or code-switching? Traces of community norms in Vietnamese-English speech. *The Australian Journal of Linguistics* 38(4), 443–466. doi:10.1080/07268602.2018.1510727. Cited on pages 2, 18, 41, 43, 46, 51, 52, 56, 99, 125, 136, 165, and 209.
- Nguyen, Li & Christopher Bryant. 2020. CanVEC - The Canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the 2020 international conference on language resources and evaluation*, 4121–4129. Marseille, France: Language Resources and Evaluation. <https://www.aclweb.org/anthology/2020.lrec-1.507/>. Cited on pages viii, 7, and 48.

- Nguyen, Thanh-Bon, Thi Minh Huyen Nguyen, Laurent Romary & Xuan Luong Vu. 2004. Developing tools and building linguistic resources for Vietnamese morpho-syntactic processing. In *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)*, Lisbonne, Portugal: Workshop on Asian Language Resources. <https://hal.inria.fr/inria-00107761/document>. Cited on page 129.
- Nguyen, Thi Thuy Minh & Ho Gia Anh Le. 2013. Requests and politeness in Vietnamese as a native language. *Pragmatics* 23(4), 685–714. doi:10.1075/prag.23.4.05ngu. Cited on page 126.
- Nguyen, Thy Tan Lan. 2012. Code choice in the Vietnamese community in Sydney. Australian National University Phd dissertation. Cited on pages 15, 135, and 136.
- O'Grady, William, Yoshie Yamashita & Sookeun Cho. 2008. Object drop in Japanese and Korean. *Journal of Language Acquisition* 15(1), 58–68. doi:10.1080/10489220701774278. Cited on page 128.
- Orozco, Rafael. 2015. Pronominal variation in Colombian Costeño Spanish. In Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 17–37. Washington, DC: Georgetown University Press. Cited on page 149.
- Osburne, Andrea G. 1996. Final cluster reduction in English L2 speech: A case study of a Vietnamese speaker. *Applied Linguistics* 17(2), 164–181. doi:10.1093/applin/17.2.164. Cited on page 85.
- Otheguy, Ricardo & Ana Cecilia Zentella. 2012. *Spanish in New York: Language contact, dialect levelling, and structural continuity*. Oxford: Oxford University Press. Cited on pages 84, 138, 176, and 184.
- Otheguy, Ricardo, Ana Celia Zentella & David Livert. 2007. Language and dialect contact in Spanish in New York: Toward the formation of a speech community. *Language* 83(4), 770–802. doi:10.1353/lan.2008.0019. Cited on pages 132, 133, 137, 138, 176, and 177.
- Owens, Jonathan, Robin Dodsworth & Mary Kohn. 2013. Subject expression and discourse embeddedness in Emirati Arabic. *Language Variation and Change* 25(2), 255–285. doi:10.1017/S0954394513000173. Cited on pages 150 and 207.
- O'Grady, William, Hye-Young Kwak, On-Soon Lee & Miseon Lee. 2011. An emergentist perspective on heritage language acquisition. *Studies in Second Language Acquisition* 33(2), 223–245. doi:10.1017/S0272263110000744. Cited on page 182.

- Packard, Jerome. 2000. *The Morphology of Chinese: A linguistic and cognitive approach*. Cambridge: Cambridge University Press. Cited on page 48.
- Paolillo, John. 2002. *Analyzing linguistic variation: Statistical models and methods*. Stanford, CA: CSLI. Cited on page 158.
- Papadopoulou, Despina, Eleni Peristeri, Evagelia Plemenou, Theodoros Marinis & Ianthi Tsimpli. 2015. Pronoun ambiguity resolution in Greek: Evidence from monolingual adults and children. *Lingua* 155, 98–120. doi:10.1016/j.lingua.2014.09.006. Cited on page 183.
- Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. Cited on page 57.
- Parafita Couto, Maria del Carmen, Margaret Deuchar & Marika Fusser. 2015. How do Welsh-English bilinguals deal with conflict? Adjective - noun order resolution. In Gerald Stell & Kofi Yakpo (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, 65–84. Berlin: de Gruyter. Cited on pages 76 and 99.
- Parafita Couto, Maria del Carmen & Marianne Gullberg. 2017. Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism* 23(2), 695–714. doi:10.1177/1367006917729543. Cited on page 76.
- Paredes, Silva & Lucia Vera. 1993. Subject omission and functional compensation: Evidence from written Brazilian Portuguese. *Language Variation and Change* 5(1), 35–49. doi:10.1017/S0954394500001381. Cited on page 149.
- Park, Seong Man & Mela Sarka. 2007. Parents' attitudes toward heritage language maintenance for their children and their efforts to help their children maintain the heritage language: A case study of Korean-Canadian immigrants. *Language Culture and Curriculum Culture and Curriculum* 20(3), 223–235. doi:10.2167/lcc337.0. Cited on page 16.
- Parks, Craig D. & Anh D. Vu. 1994. Social dilemma behavior of individuals from highly individualist and collectivist cultures. *Journal of Conflict Resolution* 38(4), 708–718. doi:10.1177/0022002794038004006. Cited on page 46.
- Patil, Z.N. 2008. Rethinking the objectives of teaching English in Asia. *The Asian EFL Journal Quarterly* 10(4), 227–240. Cited on page 85.

- Payne, Thomas. 1997. *Describing morphosyntax: A guide to field linguists*. Cambridge: Cambridge University Press. Cited on page 149.
- Perez-Cortes, Silvia. 2018. Acquiring obligatory and variable mood selection: Spanish heritage speakers and L2 learners' performance in desideratives and reported speech contexts. Rutgers University Phd dissertation. Cited on page 177.
- Pérez-Leroux, Ana Teresa, Alan Munn, Cristina Schmitt & Michelle DeIrish. 2004. Learning definite determiners: Genericity and definiteness in English and Spanish. In Barbara Beachley, Amanda Brown & Frances Conlin (eds.), *Supplementary proceedings of the 27th Boston University conference on language development (BUCLD)*, Somerville, MA: Cascadilla Press. Cited on page 114.
- Petrov, Slav, Dipanjan Das & Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the eighth international Conference on language resources and evaluation (LREC-2012)*, Istanbul, Turkey: European Language Resources Association (ELRA). <http://www.aclweb.org/anthology/L12-1115>. Cited on page 49.
- Pham, Van-Tinh. 2002. *Phép tính lược và ngữ thuộc tính lược trong tiếng việt [Ellipses and ellipsis phrases in Vietnamese]*. Hà Nội: Nhà xuất bản Khoa học. Cited on pages 126, 128, and 142.
- Phan, Trang & Eric Lander. 2015. Vietnamese and the NP/DP parameter. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 60(3), 391–415. doi:10.1017/S0008413100026268. Cited on page 127.
- Plaff, Carol. 1979. Constraints on language mixing. *Language* 55(2), 291–318. doi:10.2307/412586. Cited on page 99.
- Platt, John Talbot & Heidi K Platt. 1975. *The social significance of speech: An introduction to and workbook in sociolinguistics*. Amsterdam: North-Holland Publishing Company. Cited on page 32.
- Podesva, Robert. 2011. The California vowel shift and gay identity. *American Speech* 86(1), 32–51. Cited on page 122.
- Polinsky, Maria. 2011. Heritage languages. *Linguistics. Oxford Bibliographies Online Datasets* doi:10.1093/obo/9780199772810-0067. Cited on page 1.
- Polinsky, Maria. 2018. *Heritage languages and their speakers*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press. doi:10.1017/9781107252349. Cited on pages 1, 2, 172, 176, 177, 179, and 206.

- Polinsky, Maria & Gregory Scontras. 2020. Understanding heritage languages. *Bilingualism: Language and Cognition* 23(1), 4–20. doi:10.1017/S1366728919000245. Cited on pages 2, 177, 178, 179, 184, 206, and 210.
- Poplack, Shana. 1980. Sometimes i'll start a sentence in Spanish *y termino en español*: Toward a typology of codeswitching. *Linguistics* 18(7/8), 581–618. doi:10.1515/ling.1980.18.7-8.581. Cited on pages 51, 99, and 120.
- Poplack, Shana. 1988. Language status and language accommodation along a linguistic border. In Peter H. Lowenberg (ed.), *GURT 87: Language spread and language policy: issues, implications, and case studies*, 90–118. Washington, DC: Georgetown University Press. Cited on page 52.
- Poplack, Shana. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French project. In Ralph W. Fasold & Deborah Schiffrin (eds.), *Language change and variation*, 411–452. Amsterdam: John Benjamins Publishing Company. doi:10.1075/cilt.52.25pop. Cited on page 38.
- Poplack, Shana. 1993. Variation theory and language contact. In Dennis Preston (ed.), *American dialect research: An anthology celebrating the 100th anniversary of the American Dialect Society*, 251–268. Amsterdam: John Benjamins Publishing Company. doi:10.1075/z.68. Cited on pages 33, 38, and 123.
- Poplack, Shana. 2001. Code-switching (linguistic). In Niel Smelser & Paul Baltes (eds.), *International encyclopedia of the social and behavioral sciences*, 2062–2065. London/Oxford: Elsevier Science Ltd. Cited on page 159.
- Poplack, Shana. 2018. Categories of grammar and categories of speech: When the quest for symmetry meets inherent variability. In Naomi Shin & Daniel Erker (eds.), *Questioning theoretical primitives in linguistic inquiry (Papers in honor of Ricardo Otheguy)*. *Studies in functional and structural Linguistics*, 7–34. Amsterdam: John Benjamins Publishing Company. Cited on page 78.
- Poplack, Shana. In press. Data management @ the uOttawa sociolinguistics laboratory. In Andrea Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren Collister (eds.), *The open handbook of linguistic data management*, Cambridge, MA: MIT Press Open. Cited on page 39.
- Poplack, Shana & Stephen Levey. 2010. Contact-induced grammatical change: A cautionary tale. In Peter Auer, Jürgen Erich Schmidt & Alfred Lameli (eds.), *Language and space: An international handbook of linguistic variation, vol. 1: Theories and methods*, 391–419. Washington, DC: Georgetown University Press. Cited on pages 84 and 120.

- Poplack, Shana & Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International Journal of Bilingualism* doi:10.1177/136700699800200201. Cited on pages 88 and 120.
- Poplack, Shana, David Sankoff & Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26(1), 47–104. doi:10.1515/ling.1988.26.1.47. Cited on page 52.
- Poplack, Shana & Sali Tagliamonte. 1991. African American English in the diaspora: Evidence from old-line Nova Scotians. *Language Variation and Change* 3(3), 301–339. doi:10.1017/S0954394500000594. Cited on pages 134 and 154.
- Posio, Pekka. 2015. Subject pronoun usage in formulaic sequences: Evidence from Peninsular Spanish. In Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject pronoun expression in Spanish: A cross-dialectal perspective*, 59–78. Washington, DC: Georgetown University Press. Cited on page 147.
- Quesada, Margaret Lubbers. 2015. *The L2 acquisition of Spanish subjects*. Berlin: de Gruyter. doi:10.1515/9781614514367. Cited on page 176.
- Rampton, Ben. 1995. *Crossing: Language and ethnicity among adolescents*. London: Longman. Cited on page 122.
- Ranson, Dawn. 1991. Person marking in the wake of /s/ deletion in andalusian Spanish. *Language Variation and Change* 3(2), 133–152. doi:10.1017/S0954394500000491. Cited on page 147.
- Reuland, Eric. 2011. *Anaphora and language design*. Cambridge, MA: MIT Press. Cited on page 182.
- Rickford, John. 1997. Prior creolization of African-American vernacular English? Sociohistorical and textual evidence from the 17th and 18th centuries. *Journal of Sociolinguistics* 1(3), 315–336. doi:10.1111/1467-9481.00019. Cited on page 134.
- Rickford, John. 1998. The creole origins of African-American vernacular English: Evidence from copula absence. In John Rickford, Guy Bailey, John Baugh & Salikoko Mufwene (eds.), *African-American English: Structure, history and use*, 154–200. New York: Routledge. Cited on page 134.
- Rickford, John. 2006. Down for the count? The creole origins hypothesis of AAVE at the hands of the Ottawa circle, and their supporters. *Journal of Pidgin and Creole Languages* 21(1), 97–155. doi:10.1075/jpcl.21.1.03ric. Cited on page 155.

- Riehl, Claudia M. 2005. Code-switching in bilinguals: Impacts of mental processes and language awareness. In James Cohen, Jeff MacSwan, Kellie Rolstad & Kara T. McAlister (eds.), *ISB4: Proceedings of the fourth international symposium on bilingualism, 1945–1959*. Somerville, MA: Cascadilla Press. Cited on page 51.
- Rinke, Esther, Cristina Flores & Pilar Barbosa. 2017. Null objects in the spontaneous speech of monolingual and bilingual speakers of European Portuguese. *Probus* 30(1), 93–119. doi: 10.1515/probus-2017-0004. Cited on pages 133, 134, and 168.
- Romaine, Suzanne. 2012. The bilingual and multilingual community. In Tej K. Bhatia & William C. Ritchie. (eds.), *Handbook of bilingualism and multilingualism*, 445–465. Hoboken NJ: Wiley-Blackwell. Cited on page 20.
- Rothman, Jason. 2007. Pragmatic solutions for syntactic problems: Understanding some L2 syntactic errors in terms of pragmatic deficits. In Sergio Baauw, Frank Drijkoningen, Luisa Meroni & Manuela Pinto (eds.), *Romance languages and linguistic theory*, 299–320. Amsterdam: John Benjamins Publishing Company. Cited on page 176.
- Rothman, Jason. 2009. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism* 13(2), 155–163. doi: 10.1177/1367006909339814. Cited on page 176.
- Rubin, Edward & Jacqueline Toribio. 1996. Code-switching in generative grammar. In John Jensen & Ana Roca (eds.), *Spanish in Contact*, 203–226. Somerville, MA: Cascadilla Press. Cited on page 99.
- Sachdev, Itesh & Howard Giles. 2004. Bilingual accommodation. In Tej Bhatia & William Ritchie (eds.), *The handbook of bilingualism*, 353–378. Oxford: Blackwell. Cited on page 210.
- Salazar, Michelle. 2007. Está muy diferente a como era antes: *ser* and *estar* + adjective in New Mexico Spanish. In Kim Potowski & Richard Cameron (eds.), *Spanish in contact: Policy, social and linguistic inquiries*, 343–353. Amsterdam: John Benjamins Publishing Company. Cited on pages 179 and 180.
- Sanchez-Alonso, Sara. 2018. The cognitive sources of language change and variation: Connecting synchronic variation and diachrony in Spanish copula use. Yale University Phd dissertation. Cited on page 154.
- Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2), 189–222. <http://www.jstor.org/stable/4167071>. Cited on page 158.

- Saraceni, Mario. 2010. *The relocation of English: Shifting paradigms in a global era*. Basingstoke: Palgrave Macmillan. Cited on page 78.
- Saraceni, Mario. 2015. *World Englishes: A critical analysis*. London: Bloomsbury Academic. Cited on page 78.
- Saville-Troike, Muriel. 2003. *The ethnography of communication: An introduction*. Oxford: Blackwell Publishing. Cited on page 18.
- Schiffrin, Deborah. 1981. Tense variation in narrative. *Language* 57(1), 45–62. doi:10.2307/414286. Cited on page 85.
- Schilling-Estes, Natalie. 2004. Constructing ethnicity in interaction. *Journal of Sociolinguistics* 8(2), 163–195. doi:10.1111/j.1467-9841.2004.00257.x. Cited on page 122.
- Schmid, Monika. 2009. On L1 attrition and the linguistic system. *EUROSLA Yearbook* 9, 212–244. doi:10.1075/eurosla.9.11sch. Cited on page 171.
- Schönenberger, Manuela. 2001. *Embedded V-to-C in child grammar: The acquisition of verb placement in Swiss German*. Dordrecht: Kluwer. Cited on page 114.
- Schroter, Verena. 2019. *Null subjects in Englishes. A comparison of British English and Asian Englishes*. Berlin: de Gruyter. doi:10.1515/9783110649260. Cited on page 147.
- Schuler, Kathryn D., Charles Yang & Elissa L. Newport. 2016. Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In Anna Papafragou, Daniel J. Grodner, Daniel Mirman & John Trueswell (eds.), *Proceedings of the 38th annual conference of the cognitive science society*, 2321–2326. Philadelphia, PA: Cognitive Science Society. Cited on page 114.
- Schwenter, Scott A. 2006. Null objects across South America. In Timothy L. Face & Carol L. Klee (eds.), *Selected proceedings of the eighth Hispanic linguistics symposium*, 23–36. Somerville, MA: Cascadilla Press. Cited on page 153.
- Schwenter, Scott A. 2014. Two kinds of differential object marking in Portuguese and Spanish. In Patrícia Amaral & Ana Maria Carvalho (eds.), *Portuguese/Spanish interfaces*, 237–260. Amsterdam: John Benjamins Publishing Company. Cited on pages 133 and 153.
- Sciullo, Anne-Marie Di, Pieter Muysken & Rajendra Singh. 1986. Government and code-mixing. *Linguistics* 22(1), 1–24. <https://www.jstor.org/stable/4175815>. Cited on page 99.

- Sennrich, Rico & Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 211–221. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1021. Cited on page 209.
- Serratrice, Ludovica, Antonella Sorace, Francesca Filiaci & Michela Baldo. 2009. Bilingual children's sensitivity to specificity and genericity: Evidence from metalinguistic awareness. *Bilingualism: Language and Cognition* 12(2), 239–257. doi:10.1017/S1366728909004027. Cited on page 181.
- Sharma, Devyani & John Rickford. 2009. AAVE/creole copula absence: A critique of the imperfect learning hypothesis. *Journal of Pidgin and Creole Languages* 24(1), 53–90. doi:10.1075/jpcl.24.1.03sha. Cited on page 155.
- Shenk, Petra Scott. 2006. The interactional and syntactic importance of prosody in Spanish-English bilingual discourse. *International Journal of Bilingualism* 10(2), 179–205. doi:10.1177/13670069060100020401. Cited on pages 41 and 42.
- Silva-Corvalán, Carmen. 1986. Bilingualism and language change: The extension of *estar* in Los Angeles Spanish. *Language* 62(3), 587–608. doi:10.2307/415479. Cited on page 179.
- Silva-Corvalán, Carmen. 1994. *Language contact and change: Spanish in Los Angeles*. Oxford: Clarendon Press. Cited on pages 176 and 179.
- Silva-Corvalán, Carmen & Andrés Enrique-Arias. 2017. *Sociolingüística y pragmática del Español*. Washington, DC: Georgetown University Press. Cited on page 147.
- Singler, John. 2001. Why you can't do a VARBRUL study of quotatives and what such a study can show us. *University of Pennsylvania Working Papers in Linguistics* 7, 257–278. Cited on page 85.
- Sloetjes, Han & Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA). [http://www.lrec-conf.org/proceedings/lrec2008/pdf/208\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf). Cited on page 37.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, doi:10.1.1.129.4369. Cited on page 57.

- Sofu, Hatice. 2009. Language shift or maintenance within three generations: Examples from three Turkish-Arabic-speaking families. *International Journal of Multilingualism* 6(3), 246–257. doi:10.1080/14790710902878684. Cited on page 69.
- Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang & Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the first workshop on computational approaches to code-switching*, 62–72. Doha, Qatar: Association for Computational Linguistics. doi:10.3115/v1/W14-3907. Cited on page 48.
- Solorio, Thamar & Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 1051–1060. Honolulu, Hawaii: Association for Computational Linguistics. <http://www.aclweb.org/anthology/D08-1110>. Cited on page 48.
- Song, Chenchen. 2019. On the formal flexibility of syntactic categories. University of Cambridge Phd dissertation. Cited on page 126.
- Sorace, Antonella. 2000. Differential effects of attrition in the L1 syntax of near-native L2 speakers. In *Proceedings of the Annual Boston University Conference on Language Development*, 719–725. Boston: Cascadilla Press. Cited on pages 176 and 177.
- Sorace, Antonella. 2004. Native language attrition and developmental instability at the syntax–discourse interface: Data, interpretations and methods. *Bilingualism: Language and Cognition* 7(2), 143–145. doi:10.1017/S1366728904001543. Cited on page 176.
- Sorace, Antonella. 2011. Pinning down the concept of “interface” in bilingualism. *Linguistic Approaches to Bilingualism* 1(1), 1–33. doi:10.1075/lab.1.1.01sor. Cited on pages 4, 5, 176, 180, 183, and 205.
- Sorace, Antonella. 2016. Referring expressions and executive functions in bilingualism. In Irina Sekerina & Lauren Spradlin (eds.), *Bilingualism and executive function: An interdisciplinary approach*, 669–684. Amsterdam: John Benjamins Publishing Company. Cited on pages 4, 5, 180, and 205.
- Sorace, Antonella & Francesca Filiaci. 2006. Anaphora resolution in near-native speakers of Italian. *Second Language Research* 22(3), 339–368. doi:10.1191/0267658306sr271oa. Cited on pages vi, 4, 5, 132, 172, 175, 176, 180, 182, 203, and 205.

- Sorace, Antonella & Ludovica Serratrice. 2009a. Internal and external interfaces in bilingual language development: Beyond structural overlap. *International Journal of Bilingualism* 13(2), 195–210. doi:10.1177/1367006909339810. Cited on pages 4, 5, 180, and 205.
- Sorace, Antonella, Ludovica Serratrice, Francesca Filiaci & Michela Baldo. 2009b. Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua* 119(3), 460–477. doi:10.1016/j.lingua.2008.09.008. Cited on pages 5, 180, 181, and 205.
- Spronck, Stef & Tatiana Nikitina. 2019. Reported speech forms a dedicated syntactic domain. *Linguistic Typology* 23(1), 119–159. doi:10.1515/lingty-2019-0005. Cited on page 96.
- Stanford, James. 2016. A call for more diverse sources of data: Variationist approaches in non-English contexts. *Journal of Sociolinguistics* 20(4), 525–541. doi:10.1111/josl.12190. Cited on page 2.
- Tagliamonte, Sali. 2004. “He’s like, she’s like”: The quotative system in Canadian youth. *Journal of Sociolinguistics* 8(4), 493–514. doi:10.1111/j.1467-9841.2004.00271.x. Cited on page 85.
- Tagliamonte, Sali. 2006. *Analysing sociolinguistic variation: Key topics in sociolinguistics*. Cambridge: Cambridge University Press. Cited on page 137.
- Tagliamonte, Sali. 2007. Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change* 19(2), 199–217. doi:10.1017/S095439450707007X. Cited on pages 78 and 85.
- Tagliamonte, Sali. 2011. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley-Blackwell. Cited on pages 78, 124, and 155.
- Tagliamonte, Sali. 2012. *Analysing sociolinguistic variation. Key topics in sociolinguistics*. Cambridge: Cambridge University Press. Cited on pages 27, 78, 120, and 159.
- Tagliamonte, Sali & 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. doi:10.1017/S0954394512000129. Cited on pages 120, 123, 156, 158, and 159.
- Tao, Hongyin. 1996. *Units in Mandarin conversation: Prosody, discourse, and grammar*. Studies in discourse and grammar ; v.5. Amsterdam: John Benjamins Publishing Company. Cited on page 41.

- Thai, Bao Duy. 2005. Code choice and code convergent borrowing in Canberra Vietnamese. In Thao Le (ed.), *Proceedings of the international conference on critical discourse analysis: Theory into research*, Tasmania: University of Tasmania. Cited on page 2.
- Thieberger, Nicholas & Andrea L. Berez. 2012. *Linguistic data management*. Oxford: Oxford University Press. Cited on page 37.
- Thomas, Mandy. 1999. *Dreams in the shadows: Vietnamese-Australian lives in transition*. Sydney, NSW: Allen and Unwin. Cited on page 14.
- Thomason, Sarah & Terrence Kaufman. 1998. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press. Cited on pages 69 and 134.
- Thompson, Laurence. 1965. *A Vietnamese grammar*. Seattle, WA: University of Washington Press. Cited on page 124.
- Thompson, Lester & John Stannard. 2008. Australian values, liberal traditions and Australian democracy: Introductory considerations of government for contemporary civil society. *Social Alternatives* 27(1), 58–63. Cited on page 164.
- Timm, Lenora. 1975. Spanish-English code-switching: *El porque y how not to*. *Romance Philology* 28(1), 473–482. doi:10.1558/sols.v8i1.23. Cited on page 99.
- Ton, Thoai Nu-Linh. 2018. Ellipsis of terms of address and reference in casual communication events in Vietnamese. *Language and Linguistics* 19(1), 196–208. doi:10.1075/lali.00007.ton. Cited on page 126.
- Torres Cacoullou, Rena & Catherine E. Travis. 2015. Gauging convergence on the ground: Code-switching in the community. *International Journal of Bilingualism* 19(4), 365–480. doi:10.1177/1367006913516046. Cited on pages 34, 35, 42, and 43.
- Torres Cacoullou, Rena & Catherine E. Travis. 2018. *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge: Cambridge University Press. doi:10.1017/9781108235259. Cited on pages 21, 27, 36, 38, 39, 43, 45, 46, 51, 83, 131, 138, 147, 149, 168, 169, and 207.
- Tran, Jennie. 2011. The acquisition of Vietnamese classifiers. University of Hawai'i at Mānoa Phd dissertation. Cited on pages 109, 110, 111, 112, and 113.
- Travis, Catherine E. 2005. *Discourse markers in Colombian Spanish. A study in polysemy*. Berlin, Boston: de Gruyter. Cited on page 41.

- Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19(2), 101–135. doi:10.1017/S0954394507070081. Cited on page 149.
- 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. doi:10.1017/S0954394515000174. Cited on pages 146, 150, and 207.
- Travis, Catherine E. & Rena Torres Cacoullos. 2013. Making voices count: Corpus compilation in bilingual communities. *Australian Journal of Linguistics* 33(2), 170–194. doi:10.1080/07268602.2013.814529. Cited on page 45.
- Treffers-Daller, Jeanine. 2005. Evidence for insertional codemixing: Mixed compounds and French nominal groups in Brussels Dutch. *International Journal of Bilingualism* 9(3-4), 477–508. doi:10.1177/13670069050090030901. Cited on page 51.
- Trudgill, Peter. 1974. *Sociolinguistics: An introduction*. London: Penguin Harmondsworth. Cited on pages 32 and 155.
- Trudgill, Peter. 1986. *Dialects in contact*. Hoboken, NJ: Wiley-Blackwell. Cited on page 210.
- Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press. Cited on pages 116 and 134.
- Trueswell, John, Irina Sekerina, Nicole Hill & Marian Logrip. 1999. The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition* 73(2), 89–134. doi:10.1016/S0010-0277(99)00032-3. Cited on page 34.
- Truong, Vinh Ky. 2003. *Grammaire de la langue annamite*. Saigon: Guiland & Martinon. Cited on page 111.
- Tsimpli, Ianthi. 2014. Early, late or very late?: Timing acquisition and bilingualism. *Linguistic Approaches to Bilingualism* 4, 283–313. doi:10.1075/lab.4.3.01tsi. Cited on pages 4, 5, 180, 181, 182, 184, and 205.
- Tsimpli, Ianthi & Antonella Sorace. 2006. Differentiating interfaces: L2 performance in syntax–semantics and syntax–discourse phenomena. In *Proceedings of the 30th annual Boston University conference on language development*, 653–664. Somerville, MA: Cascadilla Press. Cited on pages 181 and 183.

- Tsimpli, Ianthi, Antonella Sorace, Caroline Heycock & Francesca Filiaci. 2003. Subjects in L1 attrition: Evidence from Greek and Italian near-native speakers of English. *Proceedings of the 27th annual Boston University conference on language development* 787–797. Cited on page 176.
- Tsimpli, Ianthi, Antonella Sorace, Caroline Heycock & Francesca Filiaci. 2004. First language attrition and syntactic subjects: A study of Greek and Italian near-native speakers of English. *International Journal of Bilingualism* 8(3), 257–277. doi:10.1177/13670069040080030601. Cited on pages 176 and 177.
- Tuc, Ho-Dac. 2003. *Vietnamese-English bilingualism: Patterns of code-switching*. London: Routledge. Cited on pages 2, 13, 28, 29, 34, 78, 135, 136, and 209.
- Unsworth, Sharon. 2013. Current issues in multilingual first language acquisition. *Annual Review of Applied Linguistics* 33(1), 21–50. doi:10.1017/S0267190513000044. Cited on page 184.
- Unsworth, Sharon. 2015. Amount of exposure as a proxy for dominance in bilingual language acquisition. In Carmen Silva-Corvalán & Jeanine Treffers-Daller (eds.), *Language dominance in bilinguals: Issues of measurement and operationalisation*, 156–173. Cambridge: Cambridge University Press. Cited on page 171.
- Valian, Virginia. 2016. Null subjects. In Jeffrey Lidz, William Snyder & Joe Pater (eds.), *The Oxford handbook of developmental linguistics*, 386–413. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199601264.013.17. Cited on pages 171 and 192.
- van Gelderen, Elly & Jeff MacSwan. 2008. Interface conditions and code-switching: Pronouns, lexical DPs, and checking theory. *Lingua* 118(6), 765–776. doi:10.1016/j.lingua.2007.05.003. Cited on page 99.
- Vu, Thi Thanh Huong. 1997. Politeness in modern Vietnamese. A sociolinguistic study of a Hanoi speech community. University of Toronto Phd dissertation. Cited on page 126.
- Vu, Thi Thanh Huong. 1999. Gián tiếp và lịch sự trong lời cầu khẩn tiếng Việt [Indirectness and politeness in Vietnamese requests]. *Ngôn ngữ* 1, 34–43. Cited on page 126.
- Wagner, Michael. 2010. Prosody and recursion in coordinate structures and beyond. *Natural Language Linguist Theory* 28, 183–237. Cited on page 96.
- Walker, James. 2000. Rephrasing the copula: Contraction and zero in early African American English. In Shana Poplack (ed.), *The English history of African American English*, 35–72. Malden, MA: Blackwell. Cited on page 154.

- Walker, James & Marjory E. Meechan. 1999. The decreolization of Canadian English: Copula contraction and prosody. In John Jensen & Gerard van Herk (eds.), *Canadian Linguistic Association annual conference proceedings 1998*, 431–441. Ottawa: Department of Linguistics, University of Ottawa. Cited on page 154.
- Walker, James & Miriam Meyerhoff. 2006. Zero copula in the Eastern Caribbean: Evidence from Bequia. *American Speech* 81(2), 146–163. doi:10.1215/00031283-2006-010. Cited on pages 154 and 155.
- Wang, Lin & Haitao Liu. 2013. Syntactic variations in Chinese–English code-switching. *Lingua* 123, 58–73. doi:10.1016/j.lingua.2012.10.003. Cited on page 83.
- 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. <http://www.jstor.org/stable/20011376>. Cited on page 171.
- Wang, Sung Lan. 2007. Evaluating competing models of code-switching with reference to Mandarin/Tsou and Mandarin/Southern Min data. University of Wales Phd dissertation. Cited on pages 5, 73, 74, 76, 77, 78, 82, 83, 84, 101, 103, 104, 105, 106, 107, and 110.
- Wang, Sung-Lan. 2016. On determining matrix language of code-switching between Southern Min and Mandarin. *The Journal of Chinese Linguistics* 44(2), 357–383. doi:10.1353/jcl.2016.0014. Cited on pages 5, 101, and 104.
- Wei, Longxing. 2000. Types of morphemes and their implications for second language morpheme acquisition. *International Journal of Bilingualism* 4(1), 29–43. doi:10.1177/13670069000040010301. Cited on page 115.
- Weinert, Regina & Jim Miller. 1996. Cleft constructions in spoken language. *Journal of Pragmatics* 2(2), 173–206. doi:10.1016/0378-2166(94)00079-4. Cited on page 96.
- Weinreich, Uriel, William Labov & Marvin Herzog. 1968. Empirical foundations for a theory of language change. In Winfred Philipp Lehmann & Yakov Malkiel (eds.), *Directions for historical linguistics: A symposium*, 95–188. Texas: University of Texas Press. Cited on pages 123 and 134.
- Weir, Andrew. 2012. Left-edge deletion in English and subject omission in diaries. *English Language and Linguistics* 16(1), 105–129. doi:10.1017/S136067431100030X. Cited on page 169.
- Wenger, Etienne. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press. doi:10.1016/j.gloenvcha.2005.04.004. Cited on pages 20 and 21.

- Wetzer, Harrie. 2013. *The typology of adjectival predication*. Berlin: de Gruyter. doi:10.1515/9783110813586. Cited on page 130.
- Winford, Donald. 2003. *An introduction to contact linguistics*. Malden, Oxford: Blackwell. Cited on page 51.
- Winford, Donald. 2009. On the unity of contact phenomena and their underlying mechanisms. In Ludmila Isurin, Donald Winford & Kees de Bot (eds.), *Multidisciplinary approaches to code-switching*, 279–306. Philadelphia: PA John Benjamins Publishing Company. Cited on page 51.
- Wolfram, Walt. 1993. Ethical considerations in language awareness programs. *Issues in Applied Linguistics* 4(2), 225–255. Cited on page 124.
- Wolfram, Walt & Caroline Myrick. 2017. Linguistic commonality in English of the African diaspora: Evidence from lesser-known varieties of English. In Cecelia Cutler, Zvezdana Vrzić & Philipp Angermeyer (eds.), *Language contact in Africa and the African diaspora in the Americas: In honor of John V. Singler*, 145–175. Amsterdam: John Benjamins Publishing Company. Cited on page 134.
- Wong, Andrew. 2005. The reappropriation of Tongzhi. *Language in Society* 34(5), 763–793. doi:10.1017/S0047404505050281. Cited on page 122.
- Wong, Cathy Sin Ping. 1987. The acquisition of Cantonese noun phrases. University of Hawaii Phd dissertation. Cited on page 111.
- Wyngaerd, E. Vanden. 2017. The adjective in Dutch-French codeswitching: Word order and agreement. *International Journal of Bilingualism* 21(4), 454–473. doi:10.1177/1367006916632302. Cited on page 76.
- Yang, Charles & Silvina Montrul. 2017. Learning datives: The tolerance principle in monolingual and bilingual acquisition. *Second Language Research* 33(1), 119–144. doi:10.1177/0267658316673686. Cited on page 114.
- Yuan, Boping. 1997. Asymmetry of null subjects and null objects in Chinese speakers' L2 English. *Studies in Second Language Acquisition* 19(4), 467–497. doi:10.1017/S0272263197004038. Cited on pages 143 and 153.
- Zhang, Liang, Aijun Li & Yingyi Luo. 2018. Chinese causal relation: Conjunction, order and focus-to-stress assignment. In *The eleventh international symposium on Chinese spoken language processing (ISCSLP)*, 339–343. Cited on page 96.

- Zhang, Qing. 2005. A Chinese yuppie in Beijing: Phonological variation and the construction of a new professional identity. *Language in Society* 34(3), 431–466. doi:10.1017/S0047404505050153. Cited on page 122.