



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

An evolutionary approach to bilingualism

Seán Geraint Roberts



A Thesis submitted for the degree of Doctor of Philosophy

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

October 2012

ABSTRACT

The ability to learn multiple languages simultaneously is a fundamental human linguistic capacity. Yet there has been little attempt to explain this in evolutionary terms. Perhaps one reason for this lack of attention is the idea that monolingualism is the default, most basic state and so needs to be explained before considering bilingualism. When thinking about bilingualism in this light, a paradox appears: Intuitively, learning two languages is harder than learning one, yet bilingualism is prevalent in the world. Previous explanations for linguistic diversity involve appeals to adaptation for group resistance to freeriders. However, the first statement of the paradox is a property of individuals, while the second part is a property of populations. This thesis shows that the properties of cultural transmission mean that the link between individual learning and population-level phenomena can be complex. A simple Bayesian model shows that just because learning one language is easier than two, it doesn't mean that monolingualism will be the most prevalent property of populations.

Although this appears to resolve the paradox, by building models of bilingual language evolution the complexity of the problem is revealed. A bilingual is typically defined as an individual with “native-like control of two languages” (Bloomfield, 1933, p. 56), but how do we define a native speaker? How do we measure proficiency? How do we define a language? How can we draw boundaries between languages that are changing over large timescales and spoken by populations with dynamic structures? This thesis argues that there is no psychological reality to the concept of discrete, monolithic, static ‘languages’ - they are epiphenomena that emerge from the way individuals use low-level linguistic features. Furthermore, dynamic social structures are what drives levels of bilingualism. This leads to a concrete definition of bilingualism: The amount of linguistic optionality that is conditioned on social variables.

However, integrating continuous variation and dynamic social structures into existing top-down models is difficult because many make monolingual assumptions. Subsequently, introducing bilingualism into these models makes them qualitatively more complicated. The assumptions that are valid for studying the general processes of cultural transmission may not be suitable for asking questions about bilingualism. I present a bottom-up model that is specifically designed to address the bilingual paradox. In this model, individuals have a general learning mechanism that conditions linguistic variation on semantic variables and social variables such as the identity of the speaker. If speaker identity is an important conditioning factor, then ‘bilingualism’ emerges. The mechanism required to learn one language in this model can also learn multiple languages. This suggests that the bilingual paradox derives from focussing on the wrong kind of question. Rather than having to explain the ability to learn multiple languages simultaneously as an adaptation, we should be asking how and why humans developed a flexible language learning mechanism.

This argument coincides with a move in the field of bilingualism away from asking ‘how are monolinguals and bilinguals different?’ to ‘how does the distribution of variation affect the way children learn?’. In this case, while studies of language evolution look at how learning biases affect linguistic variation, studies of bilingualism look at how linguistic variation affects learning biases. I suggest that the two fields have a lot to offer each other.

Gwerth cynnydd yw gwarth cenedl
a'i hedd yw ei hangau hi.

Etifeddiaeth, Gerallt Lloyd Owen

There are no handles upon a language
Whereby men take hold of it
And mark it with signs for its remembrance.
It is a river, this language,
Once in a thousand years
Breaking a new course
Changing its way to the ocean.
It is mountain effluvia
Moving to valleys
And from nation to nation
Crossing borders and mixing.
Languages die like rivers.
Words wrapped round your tongue today
And broken to shape of thought
Between your teeth and lips speaking
Now and today
Shall be faded hieroglyphics
Ten thousand years from now.
Sing - and singing - remember
Your song dies and changes
And is not here to-morrow
Any more than the wind
Blowing ten thousand years ago.

Languages, Carl Sandburg

DECLARATION

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Seán Roberts

ACKNOWLEDGEMENTS

My studies were funded by an Economic and Social Research Council (ESRC) 1+3 grant [ES/G010277/1].

In 2002, being convinced that I wanted to study literature at university, I chose my final application option at random from all the university courses in Great Britain. With eight words spoken in a lift, Professor Phillip Wadler convinced me that this random choice was the best. Ten years later, there are many, many people who have made my time at Edinburgh fantastic. It is traditional to acknowledge at least two categories of people: friends and colleagues. However, as with many other concepts in this thesis, and happily for me, categorical distinctions are difficult to maintain.

For changing my mind about the nature of things, I thank Simon Kirby, Jim Hurford, Antonella Sorace, Joseph Gafaranga, Richard Shillcock, Barbara Webb and Ragnhildur Seater. For putting up with over eight years of spuriousity I thank my supervisor Simon Kirby. For balanced views and expert advice, thanks to my supervisors Antonella Sorace and Kenny Smith. For developing my research skills, and for patience in the face of errors both computer and human, thanks to Jules Simner, Hannah Cornish and Mónica Tamariz. For discussing ideas and feedback I thank Chrissy Cuskley, Vanessa Ferdinand, Lauren Hall-Lew, Jim Hurford, Keelin Murray, Gareth Roberts, Bill Thompson, James Winters, all the readers and commentators of *Replicated Typo* and *The Adventures of Auck* and everyone in the LEC.

For letting me play music, thanks to Theo Pantelides, Fotis Katsikeris, Danny Robertson, Stuart Land, Charles Ross, Keelin Murray, Pippa Line, Mog Stapleton, Márton Sóskuthy and Bill Thompson. For distracting me with pointless games, thanks to Anthony Boden, Simon Morton, Alun Roberts, Eric Johnstone and David Bryden. Thank you for living with me Ali, Ash, Brandon, Mike, Ian, Márton, Rebecca, Alun, Rachael, Katie, Emma, Fezzie, Nick, Ashley and Kevin. For supplying me with the biological and cultural tools I needed, not only to complete this thesis, but to take advantage of all the opportunities I have been given, for their continuing support and for suggesting that I give various people, institutions and ideologies ‘hell’, I thank Mam, Dad, Gua, Gransha, Granny and my sisters Anwen and Róisín. Finally, for tolerance, moral support, physical support, meta-physical support, basic nutrition and being my toughest critic, thank you to Liz Irvine.

CONTENTS

1	Introduction	1
1.1	Introduction	1
1.2	Approaches to bilingualism	4
1.3	A concrete definition of bilingualism	8
1.4	A note on terminology	11
1.5	Structure of the thesis	11
2	The evolution of bilingualism	13
2.1	Introduction	13
2.2	Selectionist arguments	14
2.3	Cognitive Flexibility	16
2.4	A comparative approach	18
2.5	Conclusion	24
3	Languages are not discrete	26
3.1	Introduction	26
3.2	Measures of languageness	27
3.3	Subjective measures	31
3.4	Psycholinguistic measures	32
3.5	Defining languages through use	33
3.6	Mutual intelligibility	34
3.7	Typology	35
3.8	Descent	36
3.9	Statistical learning	41
3.10	Physiology	41
3.11	Function	42
3.12	Summary	44
3.13	Eliminating the concept of discrete languages	44
3.14	Requirements of a model of bilingual cultural evolution	46
3.15	Conclusion	48
4	Individual biases and population level phenomena	49
4.1	Introduction	49
4.2	Languages as species	50
4.3	Epiphenomena	51
4.4	Abstractions in models	53
4.5	Conclusion	56
5	Top-down approaches to iterated learning	57
5.1	Introduction	57
5.2	Mathematical models	58
5.3	Agent based approaches	60
5.4	Introduction to Bayesian modelling	62
5.5	Modelling cultural evolution in a rational framework	65
5.6	Learning multiple languages from multiple teachers	67
5.7	Learning the prior bias	71
5.8	Evolving the prior bias	71
5.9	The implications of bilingualism for top down models	78
5.10	Conclusion	87

6	Bayesian Bilingualism	88
6.1	Introduction	88
6.2	Model 1: Allowing multiple hypotheses	89
6.3	Model 2: Multiple learners and multiple languages	105
6.4	Conclusion	116
7	A bottom up model of bilingualism	121
7.1	Introduction	121
7.2	Bottom-up approaches	122
7.3	Model definition	124
7.4	Interpreting the model	135
7.5	Biases in the model	139
7.6	Example	142
7.7	Results: The cultural transmission of bilingualism in populations .	145
7.8	Conclusion	153
8	The implications for bilingualism	154
8.1	Introduction	154
8.2	Bilingualism as the marked condition	156
8.3	Categorical approaches to native competence	158
8.4	Categorical approach to languages	163
8.5	Categorical approach to bilingualism	165
8.6	A cultural evolution approach to bilingualism	169
8.7	Implications	177
8.8	Conclusion	183
9	Conclusion	185
9.1	An evolutionary approach to Bilingualism	185
9.2	Wider implications	187
9.3	Future work	190
9.4	Conclusion	191
	Appendices	193
	A Linear Regression	193
	B Simulating Parental Input Patterns	201
	C Iterated Regression Model Dynamics	206
	D Manuscripts	250

CHAPTER 1

INTRODUCTION

“Two languages in one head? No one can live at that speed.”

Eddie Izzard

1.1 Introduction

This thesis looks at the interaction between the fields of language evolution and bilingualism. This involves re-examining the link between the learning mechanisms of individuals and the linguistic properties of populations - what we might call ‘languages’. Studies of cultural evolution have used simplifying assumptions that undermine the importance of variation between speakers (see chapter 5). This thesis explores the implications of re-introducing a consideration of bilingualism into questions about the evolution of language and culture. The major contribution of this thesis is a concrete definition of bilingualism that is valid across sub-disciplines and different timescales.

Nativist approaches to linguistics have argued that the object of study should be the core linguistic knowledge that all humans share, separated from use, positing a genetically specified Universal Grammar that constrains the possible languages that individuals can learn (e.g. Chomsky, 1965). In order to study this, the object of study was abstracted to the ‘ideal speaker’, undermining the importance of linguistic variation (De Groot, 2010; Cook and Newson, 2007; Sorace, 2011b). Nativists assume that by studying the distribution of the properties of languages that are observed, linguists can make inferences about the properties of Universal Grammar (see Kirby et al., 2012 for discussion).

Proponents of cultural evolution have opposed the nativist view by arguing that, although there is a genetic basis for general processing biases, the structure of language we currently observe is influenced by cultural transmission (e.g. Kirby et al., 2007; Christiansen and Chater, 2008; Smith, 2009; Ferdinand and Zuidema, 2009; Kirby et al., 2012; Tamariz et al., 2012). Models of cultural transmission demonstrate that strong linguistic universals can emerge from populations of learners with weak innate biases. This makes it difficult to draw inferences about the learning biases of individuals from population-level phenomena (Kirby et al., 2007; Smith, 2009; Thompson et al., 2012; Kirby et al., 2012).

I will argue that researchers in the field of cultural evolution have chosen simplifying assumptions that are appropriate for opposing the nativist claims but may not be suitable for answering questions about bilingualism. All scientific studies require some amount of abstraction when approaching a particular problem, but I will demonstrate that models that place variation at their core can come to very different conclusions to those that add variation in later. In doing so, I will show that a proper consideration of bilingualism can inform the debate on how population-level phenomena emerge from interactions between individuals.

The basis of this argument will be the adoption of a concrete definition of bilingualism. While typical definitions of bilingualism involve native competence in more than one language (Bloomfield, 1933), I will show that the concepts that this kind of definition are based on are problematic. According to this thesis, bilingualism is defined as the amount of optionality that is conditioned on social variables. For example, if my linguistic signal changes depending on the identity of my interlocutor, then I am bilingual to some degree. I will demonstrate that it is possible to build an abstract model of cultural transmission based on this definition. While researchers in the field of bilingualism are interested in many factors such as processing (Roux and Trémoulet, 2002; Bialystok and Craik, 2010), executive control (Hernández et al., 2010; Sorace, 2011b), pragmatics (Gumperz, 1982), identity (Myers Scotton, 1983) and politics (Myers Scotton and Ury, 1977), the model will not take these factors into account. Instead, the aim is to reveal general principles of bilingual cultural evolution. A model that does take more factors into account or that attempts to fit a specific, real-world case can validly be built upon this more abstract model.

However, I will go further to argue that bilingualism should also be a fundamental aspect of models of cultural evolution in general. Some approaches have considered questions such as whether bilingualism can be stable over evolutionary time (e.g. Abrams and Strogatz, 2003), or why humans have developed the ability to learn multiple languages when this appears to be a redundant ability (e.g. Hurford, 1991; Hagen, 2008). I will argue that language acquisition is fundamentally about conditioning linguistic variation on semantic variation. If social variables explain the variation in the linguistic signal, then bilingualism will emerge. That is, a general learning mechanism that can condition linguistic variation on semantic variation can not only learn the elements of a single language, but multiple languages (see also Sternberg and Christiansen, 2006). The key element that is needed for the emergence of bilingualism is variation in social variables. This also requires allowing variation within and between individuals.

The definition of bilingualism used in this thesis is robust across different fields and different timescales. It can be used to unite the fields of language evolution and bilingualism. Traditionally, studies of language evolution might ask questions such as “How did the ability to learn language evolve?” or “How does the structure of language change over time?”. Studies of bilingualism might ask “How

do individuals acquire and use two languages simultaneously?”. However, I will argue that both fields can address the question of “How do linguistic variation, social structures and learning biases coevolve?”, and will benefit mutually from converging their research aims accordingly.

The rest of this introduction introduces the ‘bilingual paradox’ and the approach to bilingualism that this thesis uses. The end of this chapter gives a note on terminology and the structure of the thesis.

1.1.1 The bilingual paradox

The motivation for this thesis was an apparent paradox. Intuitively, learning two languages is harder than learning one, yet bilingualism is prevalent in the world (see page 5)¹. If individuals were biased towards monolingualism, then why is there linguistic diversity? As Sapir puts it:

“If all the individual variations within a dialect are being constantly levelled out to the dialect norm why should we have dialect differences at all? Ought not the individual variations of each locality, even in the absence of intercourse between them, to cancel out to the same accepted speech average?”

(Sapir, 1921/1970: 149-150, cited in Nettle (1999b)).

This thesis offers three resolutions to the bilingual paradox. First, from a traditional language evolution approach, the ability to learn multiple languages may not be so surprising. The second resolution comes from top down modelling of iterated learning and suggests that the inference in the paradox does not hold. The first part of the paradox (the difficulty of learning two languages) is a property of individuals, while the second part (bilingualism) is a property of populations. This thesis shows that the way languages are transmitted over generations in a social network means that the link between individual learning and population-level phenomena is complex. Indeed, just because learning one language is easier than two, it doesn’t mean that monolingualism will be the most prevalent property of populations.

The third resolution realises that discrete languages may not be the best unit of measurement of the complexity of the input, nor of whether an individual is bilingual. Rather, it is the number of conditioning factors that make learning

¹This will be referred to as ‘the bilingual paradox’ throughout this thesis. Although related, it should not be confused with Petitto and Kovelman’s ‘bilingual paradox’ (Petitto and Kovelman, 2003) which recognises that people are in awe of the learning abilities of children, yet worry that input from two languages will confuse them. Neither should the ‘bilingual paradox’ presented in this chapter be confused with Sorace’s ‘bilingual paradox’ (Sorace, 2011a) which notes that bilingual children appear to exhibit benefits in non-linguistic domains, but deficits in some areas of language competence.

difficult, and bilingualism is a property of linguistic use. If the linguistic input divides clearly along social variables, then bilingualism may emerge. The bottom up model suggests that the bilingual paradox is asking the wrong question. Instead, evolutionary linguists should be asking how linguistic diversity, cognitive biases and social structures coevolve.

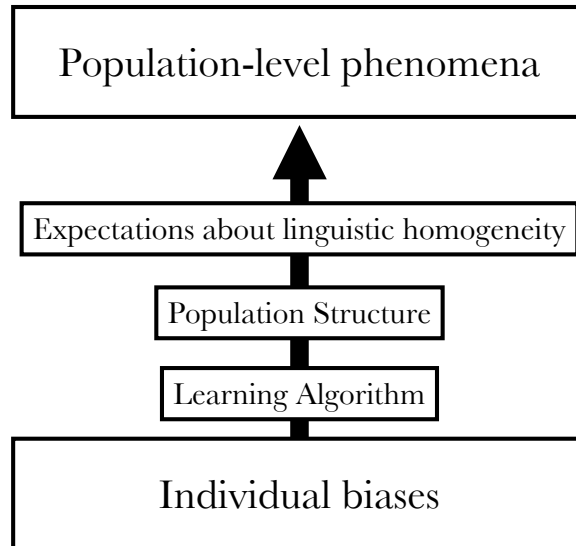


Figure 1.1: This thesis considers the link between individual learning biases and population-level cultural phenomena. This link may be complicated by factors such as the way individuals learn, the structure of the population and individual’s expectations about linguistic homogeneity or ‘bilingualism’.

1.2 Approaches to bilingualism

Bilingualism is often seen as a secondary phenomenon of investigation in linguistics (De Groot, 2010; Chomsky, 2000). Chomsky suggests that the object of study for linguistics should be the ‘ideal speaker’ who knows their language perfectly in a homogenous population (Chomsky, 1965). This approach undermines the importance of bilingualism for an evolutionary theory of the origins of language (e.g. see Chomsky, 2000; Cook and Newson, 2007). This is by no means the only approach: many other linguistic theories see variation as essential. Some ancient theories of the origins of language as far back as the 3rd century BCE emphasise the role of variation between individuals and between groups (see Nemeth, 2011). There was also opposition to the idea of idealising populations as homogenous in the early days of modern linguistics (e.g. Breal, 1893, see Andresen, 1990). There are researchers who investigate bilingualism in a generativist framework (e.g. (Roeper, 1999; Hancin Bhatt, 2008; Satterfield, 1999; MacSwan, 1999)). The debates between nativist and functionalist explanations has also become less polarised in recent times (see Hurford, 1990).

Despite these alternative approaches, the kind of nativist approach exemplified by Chomsky is taken seriously to the extent that proponents of cultural evolution often present their research as being in opposition to it². Proponents of cultural evolution also present their research as being in opposition to nativist approaches typified by Chomsky (e.g. Kirby et al., 2012). While strong nativist views assume that linguistic universals arise primarily due to biological factors (see Hauser et al., 2002; Kirby et al., 2007 for reviews), there is a growing literature that argues that linguistic universals are the product of an interaction between biological evolution, individual learning and cultural transmission, and that linking population-level phenomena to individual learning mechanisms is not straightforward (Hurford, 1990; Kirby et al., 2007; Smith, 2009; Thompson et al., 2012; Levinson and Gray, 2012; Dunn et al., 2011, see chapter 3). Much of this work has been carried out through abstract computational models of cultural transmission (e.g. Bayesian, agent-based models, phylogenetic models). While the models demonstrate a complex link between individual biases and population-level phenomena, the initial models adopted the ‘ideal speaker’ as a model of individuals. That is, individuals in the model could only speak one language, they only learned from one teacher and individuals’ knowledge of languages was categorical.

The field of language evolution is not alone in assuming that monolingualism is the default, as criticisms of other fields such as generativism (Cook and Newson, 2007), neurolinguistics (Grosjean, 1989), language acquisition (Petitto and Kovelman, 2003), psycholinguistics (De Groot, 2010, p. 342) and sociolinguistics (Otsuji and Pennycook, 2010) demonstrate. The perception of the prevalence of bilingualism is artificially low especially for native speakers of global languages such as English (Thomas and Wareing, 1999; Kostoulas-Makrakis, 2001; Dembeck and Minnaard, 2012; Luchtenberg, 2002; Yildiz, 2012). However, exposure to multiple languages is the norm in the majority of societies. Although hard data on bilingualism is sparse due to economic, political and theoretical reasons (e.g. Paliwala, 2012, and see chapter 3), the majority of people in Europe describe themselves as knowing more than one language (54%) and trilinguals are the majority in 8 countries in Europe (European Commission, 2012) - a goal that the European Commission has set for all countries in the European Union (European Commission, 2012). This is despite the fact that the levels of bilingualism in Europe have actually declined in recent years (European Commission, 2012). There is an assumption that other regions of the world have higher levels of bilingualism than Europe, with estimates varying between 50% (Wolff, 2000) and two-thirds (Crystal and Wang, 1997) being bilingual. In some countries such

²From a total of 1303 publications included in the *Language Evolution and Computation Bibliography* (<http://groups.lis.illinois.edu/amag/langev/>), the number of citations Chomsky receives is 552 according to the list of “Other relevant highly cited publications” induced from full-text references in this bibliography. Chomsky authored 6 of the top 10 publications in this list. For the 993 publication records that included a list of references, there were 799 references to Chomsky in 347 publications (35% of all publications).

as China, up to 80% of the population are estimated to be bilingual (Baker and Jones, 1998).

However, not only is bilingualism in the lay sense the norm, I argue that variation between speakers and optionality within speakers are fundamental aspects of linguistic systems. Indeed, while some researchers see the ability to learn many languages as a problem for a theory of grammatical knowledge (e.g. Roeper, 1999), others see variation as a central feature (e.g. Sorace, 2000). If a construction has optionality, there can still be agreement between people on the level of optionality. For example, some verbs only permit a particular auxiliary while others can take more than one on a graded scale of acceptability, although there is consensus in the variation used (Sorace, 2000).

The bottom-up model, presented in chapter 7, includes a general learning mechanism which conditions linguistic variation on semantic variation and a dynamic social structure that is observable by learners. The model demonstrates that a mechanism that can learn one ‘language’ can learn many, if linguistic variation can be conditioned on social variables. While other models have had to be augmented or changed to cope with bilingualism, bilingualism emerges ‘for free’ in the bottom up model by building in variation from the start (see also a similar argument by Sternberg and Christiansen, 2006; Sternberg, 2006).

1.2.1 Defining bilingualism

In order to construct this model, a concrete definition of bilingualism is needed. Mackey (2000) notes a relaxing of the definition of a bilingual over time: Bloomfield defined it as “native-like control of two languages” (Bloomfield, 1933, p. 56) (which matches a lay understanding of the term ‘bilingualism’); Haugen qualified it as the ability to produce “complete meaningful utterances in the other language” (Haugen, 1953, p. 7); Diebold, seeing problems with this, suggested bilingualism involved “contact with possible models in a second language and the ability to use these in the environment of the native language” (Diebold, 1961, p. 111); and Mackey simply stated that bilingualism was “the alternate use of two or more languages by the same individual” (Mackey, 2000, p. 22). Most definition of bilingualism, however, involve the concepts of discrete languages and level of attainment.

It is intuitively easy to encode the Bloomfieldian notion of bilingualism in an abstract model: discrete languages can be captured by discrete variables and native control can be modelled by assuming that a speaker either knows or does not know a particular language. However, both concepts are problematic. Chapter 3 argues that identifying discrete languages is not straightforward, which means that counting the number of languages an individual speaks is difficult. The problem of defining the “native speaker”, or measuring attainment has also been questioned theoretically (e.g. Escudero and Sharwood Smith, 2001; Davies,

2003; Medgyes, 1992; Hyltenstam and Abrahamsson, 2003) and empirically (e.g. Dabrowska, 2010; Street and Dabrowska, 2010; Mulder and Hulstijn, 2011, see section 8.3). Indeed, after defining bilingualism as above, Bloomfield admits “Of course, one cannot define a degree of perfection at which a good foreign speaker becomes a bilingual: the distinction is relative” (Bloomfield, 1933, p.56).

Therefore, the definition of bilingualism typically depends on two of the most controversial and possibly least well-defined concepts in linguistics - competence and ‘languages’. Models that use concepts that are not concrete (they cannot easily be mapped onto the real world) may have little real explanatory power, or may adversely affect the direction of research. In order to look at bilingualism in a cultural evolution framework with abstract models, this thesis requires a more concrete definition of bilingualism.

Certainly there are many existing definitions of types of bilingualism (for reviews, see Butler and Hakuta, 2004; De Groot, 2010). Researchers have focussed on differences in proficiency (Peal and Lambert, 1962), language organisation (Weinreich, 1953), age of first acquisition (Genesee et al., 1978), functional ability (Diebold, 1961), the effect of a second language on the first (Lambert, 1975), sociolinguistic status (Valdés and Figueroa, 1994) and sociolinguistic identity (Hamers and Blanc, 2000). There are sometimes conflicts between different approaches. For example, Grosjean notes that all bilinguals who are not ‘balanced’ fall into an intermediate category and tend to be ignored by researchers (Grosjean, 1985).

Identifying each sub-type of bilingualism has its difficulties. For example, Butler and Hakuta (2004) identify two types of problem in determining whether an individual is a balanced or dominant bilingual: How to determine proficiency in each language and how to compare proficiencies. Similarly, Mackey (2000) suggests that the measurement of bilingualism requires consideration of the degree of competence, the function of language use, and to what extent an individual alternates between languages and is able to keep them separate.

However, nowhere in either of these articles do the authors address the problem of identifying separate languages in the first place. It is often assumed that researchers know how to divide the linguistic variation in a community into relevant sub-categories that we call language. That is, even if researchers do settle on a particular definition of a sub-type, identifying a bilingual in a particular case still relies on the intuitions of the researcher. There are examples of studies that define languages through the way individuals use them (e.g. the concept of a ‘medium’, Gafaranga, 2000, 2008), and the next chapter will explore this further.

I am not trying to argue that previous research is misguided. Studies of bilingualism typically use definitions of bilingualism that are productive given the research questions. However, I am arguing that researchers’ definitions of bilin-

gualism have been fairly interest specific and context dependent, which prevents their application across research questions. As Roeper (1999) argues,

“The concept of bilingualism has never received a widely acknowledged formal definition (to my knowledge). One can even ask: should it receive a clear formal definition? Its cousins, dialects, interlanguage, foreign language, and speech register all remain important social terms, but unclear theoretical terms. Dialects, for instance, are sometimes defined as “mutually intelligible” languages, which is a valuable human and holistic characterization, but not a formal one.”

(Roeper, 1999, p.169)

Different fields will also typically consider different levels of analysis and different timescales. Sociolinguists might ask how speakers use varieties to construct and manage identity (individuals in real time). Developmental linguists might study how children acquire multiple varieties (individuals over years). Studies of language change might consider whether bilingualism is stable over time (populations over generations). Finally, evolutionary linguists might consider the evolutionary stability of the ability to learn multiple varieties (populations over evolutionary time). The goal of this thesis is to construct a definition of bilingualism that can be applied in all of these approaches.

1.3 A concrete definition of bilingualism

This section introduces the concrete definition of bilingualism developed in this thesis. Since this definition is referred to throughout the thesis, here is as good a point to introduce it as any. Although the definition might seem obvious, it was not the starting assumption of this thesis, but the result of exploring bilingualism in an evolutionary light and of building and exploring models of cultural transmission.

The definition is the following: The amount of bilingualism in a population is the amount of optionality that is conditioned on social variables. For example, optionality might be conditioned on speaker identity. If I speak differently to Mary than I do to John, then this is some amount of bilingualism. This means that bilingualism is a continuum, not a categorical concept. The level of difference could be very small or very large - I could vary in register, accent, syntax, vocabulary or use linguistic varieties that were entirely mutually incomprehensible. The level of bilingualism depends on the amount of variation that is explained by the social variables. Any observable social variable is a candidate, including aspects of an interlocutor’s identity, social context, location, time of day and so on. A method for calculating this empirically is suggested in chapter 7.

Optionality exists wherever there are linguistic forms in competition within a

speaker's utterance. That is, if the production of one variant necessarily excludes the production of another variant (including a 'null' variant with no overt form).

This defines bilingualism as a property of populations, with a bilingual community being one with a particular *distribution* of variation. While there is work on studying linguistic variation in an evolutionary framework (Greenberg, 1963; Nettle, 1998, 1999a; Nettle and Romaine, 2000; Baker, 2003; Dunbar, 2003; Roberts, 2010b; Dunn et al., 2011; Evans and Levinson, 2009), there has not been much consideration of the distribution of variation in a population. In fact, bilingualism being a property of a population opposes some views of bilingualism. For instance, Mackey (2000) states that "if language is a property of the group, bilingualism is a property of the individual". I suggest that the view of bilingualism as a property of the individual derives from the bias to study linguistic competence rather than linguistic use. If bilingualism is defined through reactions to social variables, such as interlocutor identity, then it cannot be measured in an individual in isolation³.

It is unlikely that there are any communities of moderate size that score zero on this metric of bilingualism. Even populations generally considered to be monolingual will have some amount of linguistic variation conditioned on social variables. For example, the gender of the pronoun used to refer to an interlocutor can be conditioned on the sex of the interlocutor (e.g. Hausa has a separate 2nd person pronoun for male and female referents: Siewierska, 2011; Newman, 2000). There are also statistical tendencies in the way 'monolingual' speakers use lexical items based on context and speaker identity (Altmann et al., 2011). However, any population that is considered to involve bilingualism by researchers in the field of bilingualism should always have a higher amount of bilingualism according to this measure.

There is one element of this concept that does not intuitively fit with other conceptions of bilingualism. Under the concrete definition, if every member of the population under scrutiny speaks two varieties that are recognised elsewhere as separate languages, there is no bilingualism. That is, if everyone speaks in the same way to everyone else, and if there is no social variable on which the use of either variety is conditioned (e.g. social context), then everyone is essentially speaking the same variety. There is an awareness in the sociolinguistics literature that communities that have been 'bilingual' for a long time may be qualitatively different from those that are undergoing social change (e.g. Wei and Milroy, 1995⁴). However, contentious examples of this in the real world would be very

³It would be possible to quantify the amount of variation that a single speaker conditions on social variables. However, this would still be a measure of use and would take into account the individual's proportions of interaction with other people, effectively making it a measure of a population. Since an individual's use of language is not independent from others', this measure could also only properly be understood as being *relative* to the measure of the population.

⁴"Like other researchers who have worked with immigrant communities, we are conscious of the need for a model of on-going social and linguistic change, since code-switching and

rare. Perhaps an example would be Castilian Spanish and Catalan in Barcelona, although the varieties are divided by history, politics and identity.

‘Social variables’ are factors of the social identity of the speaker or social context of the conversation, in contrast with ‘non-social variables’, which are semantic variables that are independent of these factors such as properties of the objects being described (e.g. colour). What I am labelling ‘social variables’ can be equated with what Bell (1984) identifies as factors affecting ‘stylistic’ or ‘intraspeaker’ variation. Linguistic variation is affected by linguistic factors, such as phonological effects, and extralinguistic factors. The extralinguistic factors include interspeaker and intraspeaker factors. Interspeaker factors include the differences between speakers such as class or age. For example, a speaker from an upper class might speak differently to one from a lower class. Intraspeaker factors include the identity of the addressee or the topic of conversation⁵. Indeed, the definition of bilingualism above is close to Bell’s concept of audience design as a factor in linguistic variation. For some linguistic variables, the proportion of ‘style shift’ can be measured quantitatively. For example, Coupland (1984) demonstrated that shop assistants shifted their proportions of intervocalic [t] voicing to match the proportions used in the social class of their customers. While the linguistic difference between a change in ‘style’ (e.g. choice of lexicon or vowel) is usually at a lower level from a change in ‘language’ (e.g. syntactic differences), from the perspective of this thesis, it is difficult to draw a categorical boundary between the two (see also Gumperz, 1967). Indeed, as Bell puts it, “Audience design also accounts for bilingual or bidialectal code choices. ... The monolingual depends on a linguistic variable being used differentially among speakers in the community to make it socially evaluated and available for the individual speaker to style-shift. The bilingual situation simply sharpens the process and makes it more visible” (Bell, 1984, p. 145, 158).

While this thesis focuses on social variables concerned with speaker identity, other social variables such as formality or location are also candidates for this measure. This allows this measure of bilingualism to fit the concepts such as ‘dialect’, ‘diglossia’ or ‘register’. Indeed, some see code-switching between languages and switching between registers as analysable under a single framework (Halmari and Smith, 1994). However, this means that it is possible to have ‘bilingualism’ in a population where everyone speaks identically, but uses, for instance, a different variety for formal occasions. That is, the linguistic uniformity of a population and the amount of linguistic variation in a population are independent. Everyone can agree that there’s more than one way to say something. This approach goes against certain models of cultural evolution that see uniformity and variation as opposing ends of a single continuum (e.g. Abrams and Strogatz, 2003; Steels and

language choice patterns need to be modelled very differently from those in well-established bilingual communities (Boeschoten, 1990).” (Wei and Milroy, 1995, p. 284).

⁵While Bell (1984) also labels interspeaker variables as ‘social’ factors, here I use ‘intraspeaker’ and ‘interspeaker’, to avoid confusion with my use of ‘social’ variables above.

Belpaeme, 2005;Castelló et al., 2008, see section 5.3).

However, this measure of bilingualism is still open to the biases of the researcher in a number of ways. First, the selection of social and linguistic variables to measure is not specified. There is also the question of how to compare different levels of linguistic structure (see Winters, 2011 for some possible solutions). However, as chapter 8 will show, rather than carefully selecting a few high-level features, using as many low-level features as possible is preferable and now feasible due to advances in statistical analyses (the individual differences approach, see chapter 8).

Secondly, the choice of the population to study will affect the apparent diversity, a point also made by Greenberg (1956). In computational models, it is easy to include the entire population, and to divide it according to social factors encoded in the model. Doing this in the real world might be harder, since populations migrate and change. However, the purpose of the measure in this thesis is not to measure and compare real-world populations (although this would be a feasible and interesting extension), but to explore the general processes of bilingual cultural transmission.

Although other conceptions of bilingualism have been used in abstract models, this thesis argues that they use concepts that cannot be mapped back onto the real world on theoretical grounds. The most significant shortcoming is the failure to embed the linguistic system within a dynamic social structure. Chapter 3 will show that when linguists talk about multiple languages, they are either pointing at low-level variation or social factors such as politics, history, geography, identity and so on. The definition of bilingualism used in this thesis places social structures and variation between speakers at the centre of the model.

1.4 A note on terminology

In this thesis I will follow Gafaranga (2008) and others by using ‘bilingualism’ as an umbrella term for knowledge of *more than one language*, where some would use ‘multilingualism’. In chapter 3 it will become clear that there are many different interpretations of how to define ‘languages’, and I will argue that the most valid concept of ‘bilingualism’ is as a continuous measure of how intraspeaker variation is conditioned on social variables, not a categorical measure of competence. However, I will indicate what I mean more specifically as I discuss various concepts. A basic rule of thumb is that what I mean by ‘bilingualism’ contrasts with the focus of generativist linguists on a ‘monolingual’ ideal speaker.

Part of this thesis will also argue that there is no valid way to divide linguistic variation into discrete ‘languages’ without considering complex social variables. That is, that the concept of a ‘language’ as a monolithic, discrete, static entity

is not valid in an evolutionary framework. However, eliminating this term in a study of bilingualism is very difficult (see section 3.13), especially when real-world examples are required. I would like to make clear that there is no reason to doubt that a sociolinguist, going into a specific linguistic context could usefully and validly distinguish the languages being spoken in that community. However, this usually involves complex social variables such as politics, history, geography, power, economics and identity. Rather than try to come up with a definition of bilingualism that explains how sociolinguists carve up the linguistic diversity of the world into languages, the definition this thesis uses is a minimal, concrete definition that can be applied to any population of individuals that communicate over any timescale.

1.5 Structure of the thesis

While the bulk of the thesis concentrates on cultural evolution, chapter 2 considers a more traditional language evolution approach to bilingualism. It argues that the cognitive mechanisms that allow the learning of multiple languages are evolutionarily old. Instead of a mechanism that handles bilingualism specifically, an evolutionary approach to bilingualism should look for a general learning mechanism that conditions linguistic signals on social variables. Chapter 3 argues that languages are not discrete entities that can be separated from social factors. This supports the use of the concrete measure of bilingualism and leads to some requirements for valid models of the cultural evolution of bilingualism. Chapter 4 is a brief discussion of a problem that emerges from this: if ‘languages’ are not discrete, then how can we identify bilingualism? I use arguments from biology to suggest that bilingualism can be studied as a phenomenon in its own right. Chapter 5 reviews how bilingualism is represented in previous models of cultural transmission, concluding that most models violate the requirements discussed in chapter 3. Chapter 6 presents a top-down model that explores the implications of introducing the possibility of bilingualism. Its conclusions will differ from those of the models in the previous chapter, demonstrating that the assumptions about bilingualism can affect the direction of research. However, this model, too, uses an impoverished notion of ‘bilingualism’. Chapter 7 presents a bottom-up model of cultural transmission using the concrete definition of bilingualism and adhering to the requirements set up in chapter 3. By comparing the top down and bottom up models, I demonstrate that different assumptions can lead to different solutions to the bilingual paradox. Chapter 8 discusses the implications of this approach to bilingualism for (non-evolutionary) studies of bilingualism and language acquisition. Chapter 9 provides a short conclusion.

There are three technical appendices which explain the technical details of the models. However, the implications of the thesis should be clear without understanding these. The final appendix includes some manuscripts of work written during the time spent on this thesis (Roberts, 2011; Roberts, in press).

CHAPTER 2

THE EVOLUTION OF BILINGUALISM

“Hence even the names of things were not originally due to convention, but in the several tribes under the impulse of special feelings and special presentations of sense primitive man uttered special cries. The air thus emitted was moulded by their individual feelings or sense-presentations, and differently according to the difference of the regions which the tribes inhabited. Subsequently whole tribes adopted their own special names, in order that their communications might be less ambiguous to each other and more briefly expressed.”

Epicurus, 3rd Century B.C.E.¹

“The diversity of languages arose with the building of the Tower after the Flood, for before the pride of that Tower divided human society, so that there arose a diversity of meaningful sounds, there was one language for all nations, which is called Hebrew.”

Isidore Hispalensis, 7th Century C.E.²

2.1 Introduction

This chapter uses a traditional approach to the evolution of language to determine what the the status of the ability to learn multiple languages should be. It argues that the default assumption should be that it is evolutionarily older than the ability to learn one language. This might be counterintuitive because learning multiple languages is usually thought of as being a more complex task than learning a single language. However, if we see the necessary ingredient for bilingualism as the ability to condition signals on social variables (as the measure of bilingualism presented in chapter 1 suggests), then the default assumption above is the one that makes sense. Put another way, it seems difficult to imagine a mechanism that could learn one language that couldn't also learn multiple languages. This is part of the motivation for the bottom up model in chapter 7.

¹Epicurus Letter to Herodotus in Diogenes Laertius 10.75-6 English translation from Chilton, C. W., *The Epicurean Theory of the Origin of Language. A Study of Diogenes of Oenoanda, Fragments X and XI (W)* *The American Journal of Philology*, Vol. 83, No. 2, 1962, 160; cited in Nemeth (2011).

²Isidorus Hispalensis, 1911, 9.1.1 English translation from Isidore of Seville, *The Etymologies of Isidore of Seville*, Barney, S. A. - W. J. Lewis - J. A. Beach - O. Berghof eds., Cambridge: Cambridge University Press, 2006; cited in Nemeth (2011).

Section 1 reviews previous theories of the utility of linguistic variation between individuals, but demonstrates that there has not been a lot of work on variation within individuals. Section 2 discusses whether we should assume by default a specific cognitive mechanism for dealing with multiple languages. This provides one resolution to the bilingual paradox by suggesting that the number of discrete ‘languages’ is not necessarily the best measure of the complexity of the language learning task. Rather, it is the complexity of the conditioning factors that poses a challenge to learners. Section 3 uses a comparative approach to demonstrate that other species primarily condition signals on social variables, while conditioning signals on non-social semantic variables is more difficult. This suggests that part of a bilingual ability (conditioning signals on social variables) is evolutionarily old. This motivates a view of the evolution of language which sees a consideration of bilingualism as a core component.

2.2 Selectionist arguments

There are relatively few theories of the evolution of an ability to learn multiple languages (see Sternberg and Christiansen, 2006). However, there are several theories that explain linguistic variation as adaptive at the group level, based on co-operation. Baker (2003) suggests that the ability to conceal information from other groups using an unfamiliar language could drive the creation of different languages. Dunbar (2003) notes languages diverge easily and quickly, one explanation being that different dialects make it more difficult for freeriders to take advantage of the cooperation of others by making it easy to spot outsiders (learning to speak like a native is difficult). The emergence of linguistic diversity motivated by the resistance to freeriders was demonstrated in an experiment by Gareth Roberts (Roberts, 2010b, although the social structure was an important factor, see section 7.3.2.1 in chapter 7). However, these theories rarely consider bilingualism. Having one language that is not shared by outsiders is enough to spot and resist freeriders. Bilingualism, if anything, is an advantage to potential freeriders. These theories see an ability to acquire fluency in many languages as part of the problem.

The apparent ease of children to acquire multiple languages is seen as a puzzle, due to an assumed trade-off between adaptive advantages of being able to communicate (e.g. Hagen, 2008) and investment of finite resources into learning languages (e.g. Petitto and Kovelman, 2003). It is been argued that the language capacity has clear adaptive advantages (Pinker and Bloom, 1990; Sober, 1984, Chomsky, 1982, p.18-19³, see Hurford, 1991 for a short review), either for survival or reproduction. However, it is also been argued that there is no adaptive advantage to learning multiple languages (Hagen, 2008). Hagen argues that

³The faculty of language is “highly useful and very valuable for the perpetuation of the species and so on, a capacity that has obvious selectional value”(Chomsky, 1982, 18-19).

in the early historical environment of humans, the advantages of learning a second language in adulthood would be much less than learning one in childhood, so the ability to learn languages in adulthood would not be selected for. While there are criticisms of the precise explanation⁴, Hagen points out two important factors for the adaptationist argument. First, language acquisition ability, including bilingual acquisition, is underpinned by genetic constraints (which is widely accepted, e.g. Chomsky, 1980; Hurford, 1991; Hauser et al., 2002; Dale et al., 2010). Second, the adaptiveness of bilingualism relies on the social dynamics of communities. For example, Hagen suggests that bilingualism would be adaptive in situations where groups met often enough to set up long-term trade, but not often enough to integrate. Later chapters will demonstrate that dynamic social structures are a key part of cultural transmission where bilingualism is possible and section 5.8 in chapter 5 will consider the evolution of biases for learning multiple languages based on different fitness functions of bilingual ability.

Hurford (1991) has a similar hypothesis about the ‘critical period’ effect for first language acquisition: The ease with which a language can be learned appears to change over the life history of an individual (e.g. puberty), and these changes appear to be under genetic control. If knowing a language for longer gives an individual a greater advantage, then individuals should evolve to be good at learning language early in their lifetime⁵. Although it is hard to imagine an adaptive advantage for a *diminished* ability for language learning at a later point in life (Hurford, 1991, p.172), Hurford suggests that this is a problem of perspective. Because there is a finite amount of resources that can be devoted to language acquisition, language acquisition is simply not ‘boosted’ for later life events.

These arguments typically focus on first language acquisition or adult second language acquisition, not simultaneous early bilingual acquisition. However, the assumption of a finite capacity for language leads to the puzzlement over why humans are so good at learning two languages at once (Petitto and Kovelman, 2003). That is, bilingual acquisition is often seen as redundant. However, different languages are never directly compatible (see the literature on linguistic diversity and linguistic relativity, e.g. Fausey and Boroditsky, 2008; Boroditsky et al., 2011; Lakoff, 1990; Talmy, 1985; Levinson, 2003; Evans and Levinson, 2009). While bilingual individuals may have multiple words for the same objects, there may be different connotations associated with words in each language, so there is still contrast in meaning (e.g. Bolinger, 1968; Clark, 1987; Goldberg, 1995; Croft, 2001). The bottom-up model presented later demonstrates that unconditioned (redundant) variation is indeed not stable.

⁴Hagen suggests that gaining resources through brute force would be more adaptive than trade (which would require communication), citing archaeological evidence of violence. Hagen also suggests that early humans would have belonged to small, isolated communities that did not interact often. See Hirschfeld (2008) for a critical response to the archaeological evidence. Recent genetic studies suggest that early humans had significant amounts of contact with other communities (Henn et al., 2011).

⁵However, Hurford is cautious about framing this as ‘adaptive’.

I suggest that the confusion about the redundancy of learning multiple languages involves the use of the wrong ‘unit’ for calculating linguistic capacity. Rather than the number of languages (so that bilinguals have half the capacity in each language, see Martin-Jones and Romaine, 1986), it is the number of conditioning factors that should be counted. That is, real redundancy is the number of distinct linguistic units that encode the same meaning, where meanings are distinguished by contrasts in any semantic variable - including social variables (e.g. speaker identity). This means that a bilingual language system, as defined in this thesis, is not redundant, because the extra variation encodes social meanings. In this case, one possible measure of redundancy would be the proportion of linguistic variation not accounted for by the learner’s mental language (the residual variation). The residual variation in the linguistic systems that emerge under the bottom-up model in chapter 7 tends to reduce. Again, Hurford suggests that linguistic capacity can only be measured in principle, not in practice (Hurford, 1991, p.168), but the way researchers think about this principle can affect the kinds of questions that are asked.

The resources any child has for learning language are finite, and so there must be a trade-off between learning how linguistic variation is conditioned on social and non-social factors. For example, a child cannot commit maximum resources to learning both the differences between how each parent names objects and to learning names for as many different objects as possible⁶. However, this ‘computational level’ style expression of the problem may obscure some complexities of the actual implementation. Bilingual children reach linguistic development milestones at the same pace as monolinguals (e.g. Pearson et al., 1993; Werker and Byers-Heinlein, 2008). In order to do this, it is suggested that children raised in bilingual environments have different learning strategies, for example paying attention to different cues (Healey and Skarabela, 2009; Brojde et al., 2012), being less likely to use principles such as mutual exclusivity (Byers-Heinlein and Werker, 2009a), or becoming more flexible in their learning and so are able to acquire two languages in the time monolinguals learn one (Kovács and Mehler, 2009). This flexibility demonstrates that the language learning mechanism adapts to the input during development, and that the genetic basis for learning adapts across generations to selection pressures. The effects of these two factors are easy to confound, and may interact in complex ways themselves. Cognitive flexibility may not be specific to language learning, either, since changes in brain structure have been found after learning in other domains such as physical co-ordination skills (e.g. juggling, Draganski et al., 2004; video-game playing, Green and Bave-

⁶Burkett and Griffiths (2010) describe a Bayesian model where learners have a bias over the amount of variation to expect between speakers, which resembles this kind of trade-off. This model is reviewed in chapter 5. Frank et al. (2009) also describe a Bayesian model of word learning that has a trade-off between the size of the lexicon and the number of one-to-one mappings between objects and words. However, neither model includes social variables such as the identity of speakers. Burkett and Griffiths’s model is extended in chapter 6 to include this kind of consideration.

lier, 2008 or navigating, Maguire et al., 2000⁷)

2.3 Cognitive Flexibility

Recent studies suggest that the structure of language can be shaped by cultural processes under two pressures: learnability and expressivity (e.g. Kirby et al., 2008). However, these measures may not be straightforward. The section above argued that the learnability of a linguistic system is not absolute, since learning mechanisms can adapt to the linguistic system. Therefore, a ‘bilingual’ system may not be half as ‘learnable’ as a ‘monolingual’ system. That is it may not be possible to apply a simple ‘learnability’ measure to any given linguistic system without also considering its context. With regards to expressivity, bilinguals can express themselves through codeswitching (Gafaranga, 2007, 2008), so there are exponentially more ways to express yourself with each language you learn (This point is also made by Quay, 2008 and Hoffmann, 2001a). Furthermore, the relations between languages can change over time (Hoffmann and Stavans, 2007). Again, the expressivity of a language is a function of the context of the system, and the kinds of aspect that speakers need to express.

In fact, the potential flexibility of bilinguals (or, from another perspective, the potential speciality of monolinguals) may suggest a *relaxing* of selection pressures on the learning mechanisms, rather than active selection (Ritchie and Kirby, 2007; Deacon, 2010a). That is, when the learning abilities of children with diverse types of input (monolingual and bilingual) are considered in their entirety, the salient feature appears to be a selection for *flexibility*. So, rather than asking “why is there so much linguistic variation?”, evolutionary linguists could equally ask “why is the language acquisition capacity so flexible?”. The questions are related, but point in different directions. The question about variation is perhaps based on the premise that there are strong, language-specific constraints on language learning and, as suggested above, leads to a view of monolinguals as the norm and proper object of study and bilinguals as “two monolinguals in one” (Grosjean, 1989). That is, the presumed origin of diversity is linked to factors of the individual learning mechanism. On the other hand, the question about flexibility might be answered by saying that a flexible learning capacity is clearly more adaptive than an inflexible one in certain circumstances, those circumstances being diverse and unpredictable input. The flexibility question now leads to questions about the sources of the diversity that requires this flexibility. This thesis suggests that dynamic social structures provide one source, and possibly another is the cultural transmission process itself. Therefore, two perspectives on the same problem lead to different objects of study.

⁷These examples come from Bialystok (2011), who points out that learners typically choose to engage with these other domains, so there is a certain amount of self-selection, while the exposure to multiple languages as a child is thrust upon language learners.

A similar argument is made by Pinker and Bloom: “instead of positing that there are multiple languages, leading to the evolution of a mechanism to learn the differences among them, one might posit that there is a learning mechanism, leading to the development of multiple languages” (Pinker and Bloom, 1990, p. 716). Sternberg and Christiansen (2006) suggest that this is based on the idea that learned behaviour that is selected for can come to be genetically encoded (the Baldwin effect, Baldwin, 1896; Newman, 2002), but that this pressure reduces as the behaviour is increasingly encoded, making learning easier (Hinton and Nowlan, 1987). This means that there still might be room for variation in languages after selection. However, Sternberg and Christiansen suggest that this is also the wrong perspective:

“Like Pinker & Bloom, Baker does not directly argue for a selectionist model of language differentiation as such, but gives a reason for language differentiation after selection for the linguistic ability has already taken place. What both theories are lacking, however, is an explanation for how this language system can not only accommodate language variation across groups of individuals, but also the instantiation of multiple languages within a single individual. ”

(Sternberg and Christiansen, 2006, p. 335)

Sternberg and Christiansen argue that any account of language evolution must include an explanation of the bilingual case, since it is a fundamental ability. However, rather than being selected for, Sternberg and Christiansen suggest that the bilingual ability came “for free” as a by-product of the more general learning mechanism that was selected for. In support of this, they demonstrate that a neural network trained on two languages simultaneously is able to learn both⁸. That is, a mechanism that is capable of learning one language is capable of learning many. This demonstrates that a proper consideration of bilingualism can change the approach to, questions about, and solutions to, language evolution. I will make a similar argument: Some top-down models of cultural transmission started with monolingual assumptions and added mechanisms for dealing with bilingualism (see chapter 5). The bottom-up model I present has a single, domain-general learning mechanism. While Sternberg and Christiansen (2006) show that such a mechanism can handle two varieties in the input at once, my model demonstrates that dynamic social structures can bring about the diversity

⁸A simple recurrent network was trained to predict the next word in a corpus created by two simple grammars - English and Japanese. The input nodes represented the lexical items from both languages. The input to the network included no explicit marking of which language a word belonged to. The network was trained on sequences of words from sentences in one of the languages, with a small probability of switching languages between sentences. The trained network responded differently to the two languages, although this was “local-scale language separation rather than the emergence of two completely distinct lexicons...grouping by language and part of speech gave a highly significant result, seeming to imply that the network attends to both language and part of speech, rather than primarily focusing on one.”(Sternberg and Christiansen, 2006, p. 338). For more details of the model, see Sternberg (2006)

in the input in the first place.

If a general learning mechanism can lead to bilingualism, then we might expect to see abilities analogous to bilingualism in other animals. The following section considers this possibility.

2.4 A comparative approach

How should the ability to learn multiple languages be treated in an evolutionary framework? The sections above suggest that the current default assumption about bilingualism is that before humans had an ability to learn more than one language, they had to have an ability to learn a single language. This section uses a comparative approach to evaluate this assumption. I will demonstrate that other species exhibit aspects of the capacity for bilingualism such as optionality, inter-species communication and the ability to condition signals on social variables. The latter is important for the argument of this thesis, since I argue that bilingualism is the amount of linguistic optionality that is conditioned on social variables. I argue that the default assumption about the evolution of a bilingual ability is that it is evolutionarily older than the emergence of compositional, symbolic communication.

Before looking at other species, the comparative approach to language evolution is presented. Hauser et al. (2002) approach the study of the evolution of language by considering what elements contribute towards the ‘Faculty of Language’. In the broad sense of the term, this covers all the prerequisite elements that are required for linguistic communication. This involves cognitive capacities such as acoustic string segmentation and semantic processing, but also much more basic features such as memory. That is, features of the Faculty of Language in the broad sense (FLB) are found in humans and animals. The narrow sense of the term (FLN) refers to those capacities that are involved in language *alone*. Recursive syntax has been suggested as one example (e.g. Hauser et al., 2002, although see Pinker and Jackendoff, 2005).

The comparative approach has been used to answer the question of what belongs to FLN and to FLB. Animals have been shown to be capable of a number of processes required for language, including categorical perception of speech sounds (Kuhl and Miller, 1978) and mutual exclusivity (Kaminski et al., 2004). From studies of divergent and convergent evolution of these traits, some important features have been identified. For example, many species that exhibit vocal learning have direct neural connections between the brain and vocal motors, while non-vocal learners do not (see Doupe and Kuhl, 1999).

It has been assumed that the bilingual ability emerged from a pressure to learn multiple languages (see selectionist arguments above). That is, the bilingual abil-

ity is often assumed to belong to the faculty of language in the narrow sense (FLN, see Hauser et al., 2002). For instance, top-down approaches to cultural evolution have tended to add in sensitivity to social variables as a secondary, more complex part of the learning mechanism (see chapter 5). This suggests that they assume that conditioning on non-social semantic variables was primary. However, If other species exhibit the ability to condition complex signals on social variables, then this is evidence that the mechanisms behind bilingualism developed before human language. In that case, the bilingual ability may be part of the faculty of language in the broad sense (FLB).

2.4.1 Optionality

A basic requirement of a bilingual system is optionality. That is, the ability to respond in the same way to two different signals. There are plenty of examples of optionality in animal signalling, including species with limited cognitive processing that are evolutionarily distant from humans. For example, van Wilgenburg et al. (2010) find that ants respond in the same way to more than one chemical composition. They conclude that “similar to spoken language, the chemical language of social insects contains “synonyms,” chemicals that differ in structure, but not meaning” (van Wilgenburg et al., 2010, p. 756)⁹. Another way to think about this is optionality. Just as a human may choose to say ‘teacher’ or ‘instructor’ or a bilingual may choose one language over another, ants respond in the same way to different molecules (of course, ants don’t ‘choose’ the molecules). Perhaps a closer analogy to chemical compositions is optionality in syntax. I could say “Mary gave John a book” or “Mary gave a book to John” - slightly different orders with the same basic meaning. If I change the order too much, the meaning becomes different (e.g. “John gave Mary a book”).

Another analogue of bilingualism might be the use of multimodal cues. For instance, primates communicate through vocalisations, hand gestures (Hobaiter and Byrne, 2011; Call and Tomasello, 2007; Hopkins and Leavens, 1998), facial gestures (Redican and Rosenblum, 1975) and olfactory cues (Jolly, 1966). Ay et al. (2007) demonstrate that this robustness is a property that naturally emerges in signalling systems. Furthermore, they prove mathematically that robustness is a lower bound on the complexity of a signalling system. This result is linked to the results of iterated learning experiments (e.g. Kirby et al., 2008)¹⁰.

⁹Although see Phipps, C. ‘Ant synonyms and linguistics envy’ <http://thelousylinguist.blogspot.co.uk/2010/08/ant-synonyms-and-linguistics-envy.html>

¹⁰This is not an analogy, but a formal similarity. In these experiments, participants are taught labels for a structured meaning space (e.g. objects with colour, shape and movement). They are then asked to produce the labels, given the meanings. These productions are then used as the input for the next participant in a chain of learners. If there is a bottleneck on learning, for instance the participants aren’t exposed to every label, then structure emerges in the labels. Sub-parts of the labels correlate systematically with sub-parts of the meaning (e.g. all red objects will contain the same sub-string). Kirby et al. (2008) call this ‘compositionality’, but is in fact directly related to the measure of robustness used in Ay et al. (2007). In both

Therefore, we should expect communication systems that are more complex to be more robust and therefore exhibit redundancy by, for instance, having more than one signal for a given aspect of meaning. Indeed, if we see language as a complex system (Beckner et al., 2009; Cornish et al., 2009), then we should expect to observe this kind of functional redundancy (Ay et al., 2007; Winter and Christiansen, 2012).

2.4.1.1 Inter-species semantic communication

Other evidence for optionality comes from cases of inter-species semantic communication. Many species utilise the signals of other species. In a broad sense, there appear to be universal threat signals (e.g. bright colours, size, aggression, see Morton, 1977). However, these are often innately specified rather than learned, indexically linked to physical properties and evolutionary old (see also Hinton et al., 2006). However, there are cases of species responding to more specific signals of other species.

A large proportion of songbirds mimic the calls and songs of other birds (Kelley et al., 2008)¹¹. For example, Flower (2011) demonstrates that fork-tailed drongos (a passerine bird) mimic the alarm calls of meercats in order to get food. A drongo will give the call ‘honestly’ when predators are present, but will also give the call when there are no predators so that the meercats will abandon their food for the drongo to steal. The drongos even mimic the alarm calls of other birds when doing this, meaning that their false alarms do not destabilise the apparent reliability of the signal.

Other examples include the following: sika deer use food calls of macaque monkeys as a cue for finding fruit dropped by the macaques (Koda, 2012)¹²; vervet

cases, the structure in the communication system is measured using the correlation between a distance matrix between signals and a distance matrix between meanings (although there is no structured meaning space in Ay et al., 2007). Ay et al. define robustness in terms of “insensitivity to channel occlusion through channel redundancy coupled to high information flow” (Ay et al., 2007, p. 446). Their model demonstrates that “a pressure towards an increase in robustness ... will always lead to an increase in the complexity of the message by promoting structured correlation in the channels” (Ay et al., 2007, p. 445). In the iterated learning experiment, strings linked to meanings become learnable (insensitive to channel occlusion) and expressive (high information flow), in presence of a bottleneck (channel occlusion), because the same substrings refer to the same sub-meanings in many labels (redundancy), allowing labels for previously unseen meanings to be reconstructed accurately. In Ay et al. (2007)’s model, selection for robustness naturally leads the signalling system to a trade off between ‘specificity’ and ‘generality’. Similarly, in the iterated learning case, the system trades off ‘expressiveness’ and ‘learnability’ (see also Cornish et al., 2009). However, a formal proof of the equivalence of these results is beyond the scope of this thesis. Additionally, this result is couched in terms of concurrent multimodal signalling, while bilingualism is associated with having non-correlated, independent signals.

¹¹However, in the majority of cases there is no functional explanation, rather they are likely to be learning by frequency and using calls in inappropriate contexts (Kelley et al., 2008)

¹²Although it is currently unclear whether they are responding to the food calls in particular, or just the presence of macaques (Koda, 2012).

monkeys respond to the territorial and alarm calls of superb starlings (Seyfarth and Cheney, 1990); ring-tailed lemurs respond to the alarm calls of Verreaux's sifakas (another species of lemur, Oda and Masataka, 1996); and Yellow-casqued hornbill birds respond appropriately to Diana Monkeys' alarm calls (Rainey et al., 2004).

However, many of these capacities may be innate rather than learned. For example, captive ring-tailed lemurs who had never heard the sifakas' alarm calls also responded appropriately to sifaka playbacks (Oda and Masataka, 1996). Oda and Masataka argue that they are therefore responding to shared acoustic features rather than to an associated meaning. Although most examples of inter-species communication do not involve the transference of 'concepts', some examples do show evidence for this.

Zuberbühler (2000) studied communication between Diana monkeys and Campbell's monkeys. Diana monkeys respond appropriately to Campbell's monkeys' alarm calls for leopards and eagles. Furthermore, their responses suggest they are attending to the meaning rather than the acoustic signal. If a Diana monkey hears a leopard or a leopard alarm call, it calls out loudly, but if it hears a second leopard or leopard alarm, it is quieter, presumably because of the risk of predation (the same is true of eagle alarms). Diana monkeys were primed with Campbell alarms for either leopards or eagles then probed with either the sounds that eagles or leopards make (growls and shrieks). They responded loudly to each combination, apart from where the Campbell alarm corresponded to the predator type (e.g. Campbell leopard alarm followed by a leopard sound). In these cases, the Diana monkeys were quieter, suggesting that they thought the predator was already present¹³. Zuberbühler concludes that "Diana monkeys can flexibly use and assess information derived from the communication of other species" and that "semantic understanding can be based on arbitrary signals, as it is the case for word meaning" (Zuberbühler, 2000, 717).

However, there is no current evidence to suggest that Campbell's reciprocate in their comprehension of Diana Monkey's calls. The latter issue is discussed by Magrath et al. (2009) who study the alarm call responses of 3 ecologically distinct avian species and find that responses may be reciprocal, but not necessarily symmetrical. Different species reacted to each other's alarm calls in proportion to the 'reliability' of the call as a cue to one of the listener's predators. That is, not all predators of species A are predators of species B, so the A's alarms are not always reliable for species B, and species B responds appropriately. In Magrath et al.'s study, some species responded in the same way to three different calls.

The examples above are evidence of animals responding to the signals of other species. This is analogous to bilingualism in a broad sense. For instance, Diana

¹³In this sense, the Diana monkeys are adapting their vocalisations to the listener (the predator).

monkeys seem to have the same appropriate response to two different signals. However, I suggest that it is better to see these examples above as animals who exploit cues in their environment. In species with greater cognitive capacities, these cues might extend to properties beyond the speaker's identity to tertiary objects or concepts.

2.4.2 Conditioning signals on social variables

I suggest that other species *primarily* use signals that are conditioned on social variables such as the identity of the individual producing the signal. For example, identifying conspecifics (e.g. Gottlieb, 1971) or kin (e.g. Holmes and Sherman, 1983; Holmes and Mateo, 2007), finding a mate (Zahavi, 1975; Hamilton and Zuk, 1982; Evans and Hatchwell, 1992b; Candolin, 2003) or evaluating a competitor's relative strength (Rohwer, 1982; Evans and Hatchwell, 1992a; Tibbetts and Dale, 2004). It is the conditioning on other semantic variables (triadic reference) that is difficult (e.g. Tomasello, 2006; Cartmill and Byrne, 2010; Hurford, 2010; Leavens and Bard, 2011).

Although other species do condition signals on non-social semantic variables, for instance alarm calls for predators (e.g. Seyfarth et al., 1980) or the quality of food (see Slocombe et al., 2010; Slocombe and Zuberbühler, 2006), these instances tend to be restricted to species with higher cognitive capacities (see Hurford, 2010, although a counter-example would be honey bees communicating the location of nectar, von Frisch, 1947; Gould, 1974). Even when there is communication about other things, it is often combined with social meanings, for example the reliability of an alarm call is assessed by the identity of the caller (e.g. Cheney and Seyfarth, 1988). Primate vocalisations may also be modified by the social context, for instance who is in the audience (Hauser, 1993; Slocombe and Zuberbühler, 2007; Call and Tomasello, 2007).

This might suggest that the basis of a 'bilingual' ability, in the sense of the ability to condition signals on properties of the signaller, is evolutionary older than the ability to condition signals on non-social semantic features. Indeed, recent research suggests that there is a selective pressure on signalling systems of finches for conspecific identification (Deacon, 2010a; Kagawa et al., 2012).

This thesis suggests that the only evolutionarily valid way of conceptualising bilingualism is as an ability to condition signals on social variables. In this case, the answer to whether animals have this ability may be trivially 'yes'.

2.4.3 Developmental plasticity

The sections above showed that many species condition signals on social variables, but not non-social variables. Why is this the case? One answer is that

developmental plasticity is needed in order to allow more complex signalling systems. Studies of birdsong suggest that domestication can foster developmental plasticity (see below).

The domesticated Bengalese finch exhibits very complex song patterns in comparison to its wild relative the white backed munia (Okanoya, 2004). Okanoya suggests that the survival pressures were relaxed for the domesticated finch, allowing learning mechanisms to become more flexible and allow greater variation and complexity in songs. Okanoya (2010) also shows that Benglaese finches, unlike the munia, learn from many tutors, splicing whole segments of songs from many individuals to create their own song. This might introduce more complexity into the signal over time. Furthermore, Soma et al. (2009) find that chicks select tutors based on their song complexity. In this sense, the learner’s song is conditioned on social factors: the perceived fitness of the singer. Similarly, Hultsch and Todt (1989) and Hultsch and Todt (1996) report a ‘context effect’ in nightingales, whereby they separate songs learned from different tutors.

However, Kagawa et al. (2012) has shown that vocal learners who co-inhabit areas with other species of vocal learners have less complex song. That is, reliable species identification may be orthogonal to song complexity. This suggests that the development of an ability to condition signals on social variables does not necessarily lead to greater signal complexity. Only when the environmental pressures are relaxed to allow cognitive plasticity may the two aspects coevolve.

2.4.4 Bilingualism before symbolic communication

Given the observations above, a view of the evolution of semantic systems is suggested. This view is speculative, but gives an idea of the impact that thinking about ‘bilingualism’ as an evolutionarily old ability can have on theories of language evolution.

The view is something like the following. Much of animal communication is limited to and *grounded in* information relevant to shared survival interests, that is, food, predators and mating. Humans are capable of communicating about topics beyond their immediate survival needs. This difference possibly requires the ‘ungrounding’ of signals from the domains in which they evolved and ‘re-grounding’ in other domains (e.g. Flack and De Waal, 2007; Núñez, 2010). If the ability to condition signals on *individuals* could be ungrounded to allow learning from *contexts*, then this would allow a semantic system to develop. For instance, Flack and De Waal (2007) observe that a submission gesture used during conflict (an immediate response) between pigtailed macaques is also occasionally used in peaceful interactions to signal subordination (an ongoing pattern of behaviour). The submission gesture has been ‘ungrounded’ from its original context in order to signal a different meaning. A ‘bilingual’ ability to condition signals on context could have preceded a symbolic signalling system. Okanoya and Merker (2007)

have a similar hypothesis which sees string segmentation and context segmentation as necessary preadaptations for a semantic system.

This ungrounding process would require two things. First, the selection pressure on the original system needs to be lifted by some other mechanism such as a change in the environment (e.g. Isbell and Young, 1996) or domestication (e.g. Okanoya, 2004; Deacon, 2010b). Secondly, there would have to be systematic variation between groups of speakers so that generalisations could be drawn. One source of this could be contact between groups, suggesting that dynamic social structures were an important part of the development of semantic systems (see also Dunbar, 1993).

2.5 Conclusion

Asking whether non-human species have capacities for bilingualism in the broad sense may affect the way we approach bilingualism. This chapter has reviewed studies that show that animals have capacities compatible with certain aspects of bilingualism, but without other features of human language. These capacities stem from very basic abilities to respond appropriately to cues in the environment. Many other species condition signals on properties of the speaker (e.g. fitness, threat, see Gottlieb, 1971; Rohwer, 1982, see section 2.4), while conditioning signals on non-social semantic variables (e.g. food or predators) is rarer (c.f. Seyfarth et al., 1980; Slocombe et al., 2010; Slocombe and Zuberbühler, 2006).

The evolutionary picture is therefore the following. The ability to condition signals on semantic variables is evolutionarily old. Relaxed selection (e.g. due to domestication, Ritchie and Kirby, 2005; Hare et al., 2005) leads to developmental plasticity (Deacon, 2010b; Okanoya, 2004; Takahasi and Okanoya, 2010). Indeed, we see evidence of this plasticity in studies of bilingualism that show that learners adapt their learning strategies to the input (Byers-Heinlein and Werker, 2009a; Brojde et al., 2012; Healey and Skarabela, 2009; Kovács and Mehler, 2009, see section 5.9.3). Plasticity in learning allows linguistic variation which, coupled with dynamic social structures, can be conditioned on social identities. This would bring about what this thesis recognises as bilingualism.

The hypothesis here is speculative, and there is much to be added from the neuroscience and development literature. However, I suggest that it is the right default assumption for an evolutionary approach to bilingualism. Rather than being a secondary ability that should be studied after the main ‘problem’ of language evolution is solved, it can be seen as a central part of the story. Furthermore, instead of looking for a mechanism that specifically deals with the learning of multiple languages, we should be looking for a general, flexible learning mechanism that allows language learners to pick out relevant dimensions of variation

in their input. This suggests that the ‘bilingual paradox’ presented in chapter 1 is using the wrong unit of analysis. Learning two ‘languages’ is not necessarily hard, it depends on the complexity of the conditioning factors. There may be a trade-off between resources devoted to learning how linguistic variation is conditioned on social and non-social variables. The next chapter considers the psychological reality of discrete languages further.

This approach can have an impact on the way cultural evolution is modelled, and the conclusions that come out of those models. In contrast to the approach suggested in this chapter, chapters 5 and 6 discuss top down models of cultural evolution that assume that individuals have a specific mechanism for learning multiple languages. By comparing these models with a bottom up model which uses a general, flexible learning mechanism, I will argue that the top down approach can draw misleading conclusions about bilingualism.

CHAPTER 3

LANGUAGES ARE NOT DISCRETE

“I was asked by the teacher one day to identify a picture of what I knew perfectly well my mother referred to as a “spatula.” But for the life of me I could not think of the word in English.”

Phillip Roth (1969) p. 107

3.1 Introduction

The main claim of this chapter is that there is no valid way to categorise linguistic variation into languages on a purely linguistic basis. Sober (1980) calls this the ‘line-drawing’ problem¹. Put another way, when pointing to a ‘language’, linguists are either pointing to low-level linguistic features, or a mix of linguistic and non-linguistic cultural traits. I use studies of bilingualism to support this point.

If ‘languages’ are not concrete concepts, but instead epiphenomena that emerge from populations of individuals, then it might be invalid to assume they have a psychological reality in the learning mechanisms of individuals. This has implications for models of cultural evolution. The conclusion involves three requirements for such models.

The following sections will show that purely linguistic measures fail to demarcate a consistent concept of a language. Furthermore, these failures will repeatedly demonstrate the importance of low-level linguistic features, valid representation of individuals and complex, dynamic social structures. Many of the arguments below will draw on studies of bilingualism and are intended to flesh out the important features of an evolutionary theory of bilingualism.

The idea that languages are not concrete concepts is not new (see Haugen, 2009 for a history of the concepts of a ‘language’ and a ‘dialect’ and a review of different definitions of the two). However, even very recent publications feel the need to remind readers that languages are not discrete. For instance, Croft (in press) remarks that:

¹Sober borrows this from the field of biology, but applies it specifically to language: “The problem of making the idea of languages as social entities scientifically respectable seems to be beset by line-drawing problems.”(Sober, 1980, p. 396)

“... a language as a population is neither homogeneous or sharply delineated. It is a complex system of multiple, partially overlapping populations of linguistic structures of different degrees of community inclusiveness. Each individual has a linguistic repertoire which reflects her knowledge and exposure to the communities in the society in which she engages in joint actions (shared practice). One consequence of this complex population structure is that an individual is familiar with linguistic variants for (roughly) the same meanings from different communities in which she interacts. Hence a speaker has multiple variants available to her, which are associated to greater or lesser extent with specific communities in her mind. Those variants may belong to what linguists would call different languages, in a multilingual society, even though they may be combined into a single utterance (what sociolinguists call code-switching).”

(Croft, in press, sec. 4.1)

Neither are these groupings of variants stable over time. For example, Lamb (2011) discusses complex shifts in the perception of divisions between regional varieties in Scotland within the last 100 years. Even Chomsky, who advocated the study of the ‘ideal speaker’ in a homogenous community (Chomsky, 1965, p. 3) recognises that the definition of a language is dependent on many extra-linguistic factors:

“To say that people speak different languages is a bit like saying they live in different places or look different, notions that are perfectly useful for ordinary life, but are highly interest-relative. We say that a person speaks several languages, rather than several varieties of one, if the differences matter for some purpose or interest.”

(Chomsky, 2000, p. 43-44)

While it is intuitive that there exist high-level categories of linguistic features like languages (see Goldberg, 2009; Gelman, 2003), there is little controversy over the idea that the way linguistic variation is divided into languages is dependent on non-linguistic factors. However, it is worth setting out a careful argument against the idea that a ‘language’ is an abstract concept with a concrete analogue that is useful for an evolutionary approach to bilingual cultural transmission.

3.2 Measures of languageness

The concept of a language is central to the comparative approach to linguistics (see Lehmann, 1993). It is assumed that the variation in speakers’ speech can be divided into high-level categories called ‘languages’ (e.g. English, German, Pirahã) and that being able to speak two or more languages (bilingualism) is seen as qualitatively different to having optionality within a language (see Bloomfield, 1933; Grosjean, 1989). Models of cultural evolution have considered how the distribution of languages with different properties change over time (e.g. Abrams

and Strogatz, 2003; Griffiths and Kalish, 2007, see chapter 5). This is done by explicitly specifying languages in the model as monolithic objects that can be acquired directly and are fixed over time. In this chapter I will argue that there is no way to validly categorise linguistic variance into ‘languages’ on a purely linguistic basis. What linguists identify as languages are emergent groupings of lower-level variation that are dynamic and context sensitive. Of course, linguists do productively categorise linguistic variance into languages, and I am not arguing that linguists are unaware of the problems with doing so. My argument only applies to the best way of allowing a consideration of bilingualism in an evolutionary framework.

An intuitive approach to this argument is to ask whether there is any coherent way of measuring how many languages someone speaks. If there is no valid way to demarcate a consistent concept of a ‘language’, then the concept may be invalid. This would suggest that the concept of a monolithic, discrete, static language is not a concrete property that should be directly represented in a model.

Hurford (1991) argues that measuring linguistic abilities is valid in theory, although not practically:

“Of course, it is quite beyond present-day linguistics to assign actual numbers to, say, my command of English. But the central psychological realist assumption of modern (generative) linguistics is that language users enjoy potentially infinite use of finite sets of representations stored in their finite brains. If these representations, or mental grammars, are finite, then in principle they can be assigned actual numbers indicating the amount of information they contain, even though in practice the determination of what the numbers should be is out of the question.”²

(Hurford, 1991, p.173)

However, even if the number of languages a person can speak can be objectively determined, but is defined through more appropriate lower-level concepts (the ‘representations’ in the quote above), then we must question the usefulness of the high-level concept. This is essentially an eliminativist approach³.

The difference between languages has been defined using subjective measures, language use, mutual intelligibility, typology, descent, physiological measures and

²The conception of language capacity used in Hurford (1991) is an absolute scale whereby individuals with a higher capacity are more likely to survive or reproduce. However, there is no inclusion of the effects of the similarities *between* individuals (so that an individual with a small capacity in variety A has a higher chance of mating with another individual who knows A than an individual with a high capacity in variety B). That is, this conception assumes a monolingual society.

³The approach in this chapter owes a lot to Irvine (2011) which applies eliminativism to the concept of ‘consciousness’.

functional measures. I will show that these measures fail to pick out a concrete feature in the real world that is relevant for an evolutionary approach to bilingualism. First, a brief discussion measuring linguistic diversity, since measures of linguistic diversity have been used as proxies for bilingualism.

3.2.1 Measures of linguistic diversity

While attempts have been made to measure linguistic diversity, these have been biased towards capturing the diversity between individuals, not within individuals. For example, the Greenberg diversity index, as formalised by Lieberman and Dil (1981) measures the probability of any two randomly chosen people having the same mother tongue. This has been used as a canonical measure in the explanation of linguistic diversity (e.g. Nettle, 1999a). However, the Greenberg diversity index is weak when it comes to another measurement of diversity: bilingualism. The Greenberg diversity index is defined as:

$$GDI = 1 - \sum (P_i)^2 \quad (3.1)$$

Where P_i is the proportion of the total population that comprises the i^{th} language group. So, if you have a country with two languages, and each is spoken by half the population, you have a diversity of 0.5. However, an assumption is that the percentages of the population that speak a language sum to 1. If you have a country with two languages where everybody speaks both, then the GDI comes out as -1. Put another way, in a country where there are two languages, A is spoken by 75% and B is spoken by 50% (so 25% are bilingual), this yields a diversity of 0.1875. If the goal of the measure is to correctly predict the chances of any two people speaking the same language, then this value should be 75%. As bilingualism is a common part of most peoples lives, the GDI probably under-predicts diversity.

The original paper on measuring linguistic diversity (Greenberg, 1956) actually discusses several ways of calculating linguistic diversity in a population. The ‘split personality’ method counts bilinguals as multiple people - one for each of their languages. The ‘random speaker’ method calculates the probability of two people picking the same language at random from the languages they know. Both of these methods also have weighted versions which take into account the relative resemblance of each language pair. The ‘random speaker-hearer’ method calculates the probability of a hearer understanding another speaker speak a language chosen randomly from the ones they know. The ‘index of communication’ is the probability that any two people have a language in common. Greenberg is aware of the difficulties of applying these measures accurately:

“It is also to be noted that in the measures involving polylingualism described here, no account is taken of an individual’s relative command of the several languages he knows. Satisfactory measures have

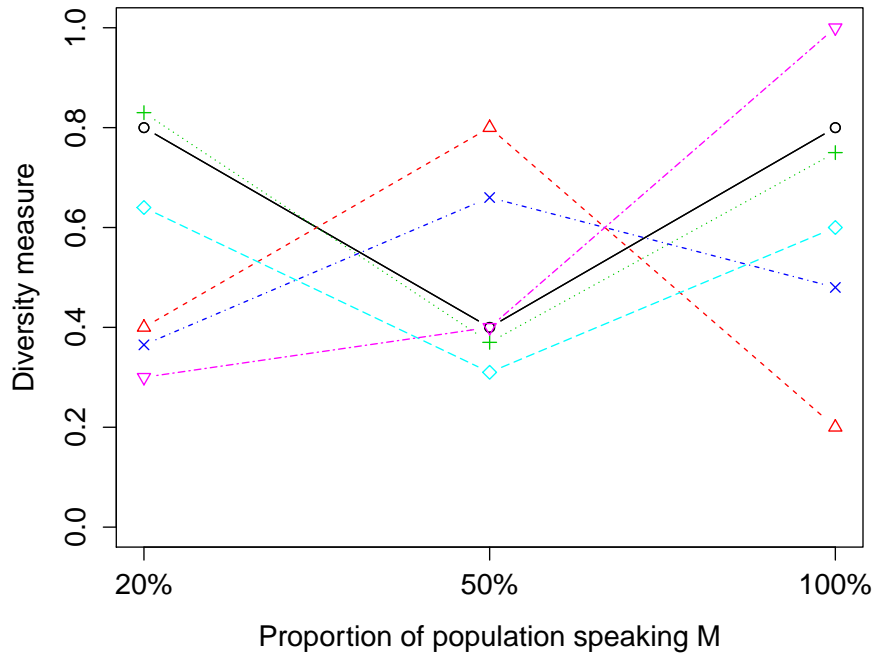


Figure 3.1: Different measures of diversity applied to the same populations from Greenberg (1956). The populations include speakers of 5 languages. In the 20% condition, each language is spoken monolingually by an equal proportion of the population. In the 50% condition, one of the languages, M, is now spoken by half of each language community. In the 100% condition, everyone in the population speaks language M. Only one type of measure changes monotonically with the prevalence of language M. Languages are related with a weighting of 0.5. The different points show the values for the different measures of diversity: split personality (black circles); weighted split personality (red triangles); random speaker (green crosses); weighted random speaker (blue xs); random speaker-hearer (blue diamonds); communication index (pink inverted triangles).

not as yet been developed, and, if developed, could hardly be applied on a scale that would allow them to be ascertained for an entire population. On the other hand, ranking as first language - usually the mother language as opposed to others - is possible, and is even found in some census reports; but there does not seem to be any non-arbitrary way of giving this fact mathematical expression”

(Greenberg, 1956, p.112)

Greenberg goes on to argue that these measures are very susceptible to the biases of the researcher, and must be considered along with the specific context in which they are applied. Figure 3.1, adapted from Greenberg (1956), demonstrates that the different methods for measuring diversity lead to very different diversity values for the same populations. Therefore, there are many possible approaches to measuring diversity. Similarly, chapter 6 demonstrates that there are many ways to measure bilingualism in top-down models.

3.3 Subjective measures

The following sections evaluate different ways of counting languages. Perhaps the most obvious way to identify the number of languages someone speaks is to assume that speakers have a good intuitive idea of the answer. Certainly, most people’s self-reports will be consistent and confident (see Blanche and Merino, 1989). However, speakers’ own intuitions about linguistic phenomena may reflect the linguistic reality poorly (see Ready-Morfitt, 1991; MacIntyre et al., 1997; and Language Log’s archive on ‘Ignorance of linguistics’ <http://languagelog.ldc.upenn.edu/n11/?cat=61>). While social groups are often identified by their language⁴, conceptions of what constitutes a language are often affected by politics, history and national identity (e.g. MacEachern, 2001; Mufwene, 2005; Haugen, 2009; Lamb, 2011) . For example, Mufwene (2005) points out that many names of languages simply refer to the way a certain (historical) group of people spoke (e.g. German means the way people called Germans speak). Particularly, the distinction between a dialect and a language may be affected by the strength of local identity (Haugen, 2009; Lamb, 2011).

Some definitions of what it means to speak a language have been offered. For instance, Haugen (1953) suggested that a minimal requirement for classifying an individual as being able to speak a language was the ability of that individual to “produce complete meaningful utterances” in that language. While this might have worked well for a particular case study, there are obvious problems with this kind of definition: What counts as ‘complete’ or ‘meaningful’? For instance, is

⁴From (Hill, 1978, p. 7): “Local groups are often named for lexical items in their dialects, e.g. *Pitjantjara*: “those having the word *pitja* ‘come’ ”; *Ngatatjara*: “those having the word *ngatja* ‘this’ ” (Gould 1969:63), and a large complex of local groups named after words for ‘yes’ and ‘no’. Another common pattern is tribal names translatable as “good speakers” or “bad speakers” (Tindale 1974:43).”

the phrase “Bonjour” a complete meaningful utterance of French and, if so, does knowing this word imply the individual speaks French? The problem of applying this definition is captured in Diebold (1961)’s description of linguistic fieldwork into bilingualism:

“...the ability to produce ‘complete meaningful utterances’ in Spanish offered a pragmatically valid boundary between bilingual and monolingual which withstood random retestings. Nevertheless, I soon became dissatisfied with writing off 81% of the San Matenios as ‘monolingual’ on the basis of Haugen’s definition of bilingual proficiency, since I had noted a minimal use of Spanish among many of these purportedly monolingual Huave-speakers. Since bilinguals are recruited from this group, it appeared that this would represent the minimal incipient situation of bilingual learning. Yet the fact remained that these individuals could not sustain even limited conversation with a Spanish-speaker, and I remained at a loss for a means of quantifying their knowledge of Spanish.”

(Diebold, 1961, p. 110)

There seems to be a trend in studies of bilingualism to assume that there is a dissociation between languages, and then to try to find a way of measuring that difference. In the case above, the subjective measure fails because the linguistic behaviour of the speakers does not fit into a simple binary category. I will discuss this problem further on, but for now let us focus on other problems with subjective measures.

Another problem pointed at here is the difficulty of measuring proficiency. As Hurford points out, “It is not actually feasible to quantify and represent in a graph the language abilities of individuals in a population, with the level of language ability on the y-axis, and numbers of individuals with these levels on the x-axis” (Hurford, 1991, p.168). Furthermore, people’s self-rating of their linguistic competence, often used in surveys, is affected by non-linguistics factors such as anxiety or perception of cultural groups (see MacIntyre et al., 1997, and section 8.3).

Even with what are traditionally thought of as well-defined languages, categorisations of linguistic variants can be very subjective (see the discussion of Thomas and Allport, 2000 in chapter 8, section 8.4 for an example). Subjective measures fail to demarcate a concrete, abstract concept of a language. Indeed, their very subjectivity suggests that languages are complex constructions which involve more than linguistic features. If this is the case, then subjective measures do not support the representation of languages as monolithic entities in a model of language evolution. In the following sections, I will show that objective measures also fail to circumscribe a coherent concept of ‘language’ that is relevant for investigating language evolution.

3.4 Psycholinguistic measures

Psycholinguistic experiments demonstrate that listeners' perceptions are different when in different language modes. This seems to suggest that discrete languages are a cognitive reality. For example, bilinguals shift their perceptual boundaries between sounds depending on which language they believe they are listening to (Elman et al., 1977; Janson and Schulman, 1983). This affect has also been demonstrated for monolinguals with knowledge of more than one accent (Niedzielski, 1999). While this point might be mitigated by allowing a broader definition of a language, perception can be altered by very subtle social cues. For example, Drager and Hay (2006) and Hay and Drager (2010) ran an experiment where speakers of English from New Zealand were presented with tokens of vowels ranging from raised and fronted (used in Australian English) to lowered and centralised (used in New Zealand English). Participants' perception of the vowels was affected by whether they had seen a stuffed toy koala (associated with Australia) or a toy kiwi bird (associated with New Zealand) prior to hearing them. As Hay and Drager suggest, linguistic and social information are closely related (see also their paper on 'sociophonetics', Hay and Drager, 2007).

If perception can be shifted by social cues, then it might be simpler to assume that languages are social concepts. At any rate, it seems simpler to assume that a single mechanism is responsible for the effects of linguistic and non-linguistic cues. Indeed, social cues can be stronger than linguistic cues (Kang and Rubin, 2009; Lindemann, 2002). Other studies found evidence of 'reverse linguistic stereotyping' as participants perceived more accented speech when they believed they were listening to a non-native speaker (Rubin, 2011; Kang and Rubin, 2009; Rubin et al., 1997)⁵. Under some conditions, semantic comprehension was actually poorer when listeners believed the speaker was non-native. This would suggest that it is better to see categories of 'languages' as being indexes of social categories. This kind of approach is the basis for the bottom-up model presented in chapter 7.

3.5 Defining languages through use

Although formal approaches might fail to demarcate a consistent concept of a 'language', speakers do use linguistic variation systematically (e.g. Auer, 1999; Gumperz and Hernandez-Chavez, 1972; Poplack, 1988; Gafaranga, 2008), so an analysis of use might provide a good measure of language. For instance, Gafaranga (2000) uses discourse analysis to analyse speakers' conceptions of language. Below is an extract of a translation of a conversation between two speakers who speak French, English and Kinyarwanda. The parts of the conversation originally in Kinyarwanda are in normal font, French is italicised and English is

⁵For an experiment looking at foreigner-directed speech and cultural evolution, see Little (2011)

in bold.

A: Refugees like him are (.) schools here are *private* (.) they are *private* so that he must pay to study at this *university*.

B: umh

A: But as he doesn't have money he has had to apply for a *grant* from the (.) they call it **local government**

B: umh

A: **Local authority** well it is like

B: // it is like a *municipality*

A: that's right it is like a *municipality* (.) he got a *grant* from the local *municipality*

(Gafaranga, 2000, p. 338)

Halfway through the conversation, speaker A forgets the French word 'municipality' and initiates a repair sequence. Since the speakers are bilingual, speaker A uses the English word to request help from speaker B to remember it. Speaker A has been switching happily between French and Kinyarwanda, so why not just use the English word to continue the conversation? Gafaranga suggests that this is evidence of a dissociation in the minds of the speakers between a Kinyarwanda-French 'medium' and an English 'medium'. That is, the speakers are treating French and Kinyarwanda as a single 'language' and English as a separate one.

Although this is a productive way of defining languages, it still does not provide a solid basis for including an abstract representation at the language level in a model of language evolution. Gafaranga's main point is that high-level categorisations of linguistic variation (what he calls 'mediums') are dynamic and context-specific. The exact partitions in variation that speakers use may change over time, and one group of speakers may partition the variation differently to another (see also Otsuji and Pennycook, 2010; Kostoulas-Makrakis et al., 2006). For example, French is considered a stand-alone 'medium' in some communities. Given this, it seems invalid to specify *a priori* fixed, high-level 'languages' in a model. Rather, the model should encode low-level variation which may be grouped by speakers into categories that can be *measured* in the model (see also Blommaert, 2010).

3.6 Mutual intelligibility

Mutual intelligibility has been used to define boundaries between varieties (see Voegelin and Harris, 1951; Yamagiwa, 1967). Simply put, if two speakers cannot understand each other, they must be speaking different languages. Nettle (1999a) demonstrates the problems with this approach. First of all, comprehension is perhaps a gradient measure and it is unclear how best to test this. For instance, should one control for pragmatic cues, a common ancestor language or cultural differences? People are adept at communicating without previous experience of

each other's languages (Levinson, 2006), and some instances of mutual intelligibility are often misleading (for example, a study of mutually intelligible words between Ainu, spoken in Japan, and Welsh: Batchelor, 1905, 1.X). Linguists can also differ in their analyses of intelligibility. For instance, Warner (1937) claimed the differences between the dialects of the Murngin from Australia were minimal and mainly "imagined", while Brendt and Brendt (1964) claimed that the dialects were actually mutually unintelligible (cited in Hill, 1978).

Secondly, even if a good measure of comprehension could be agreed upon, there is the problem of dialect chains. This is a case where a number of communities extend over a large geographic space and each community can understand its immediate neighbours, but the variation drifts until communities at extreme ends of the chain cannot understand each other. In this case, it becomes unclear where to draw the line between languages.

For example, there have been several analyses of the varieties spoken in Japan (see Shibatani, 1990, p. xiii). They have been classified based on grammatical differences into 2 dialects (Shimmura, 1904) and 3 dialects, (Tôjô, 1954) and based on phonetic similarities into a different 3 dialects (Kindaichi and Hirano, 1989). The definition of dialects are often politically motivated: when Korea was annexed by Japan, Korean was classified by a respected linguist as a 'dialect' of Japanese (see Lee and Ramsey, 2000, p.131). Many of these studies, however, realise that the boundaries are fuzzy and serve a particular purpose (Tôjô, 1938, see Preston and Long, 2002, 177-178). Tokugawa and Miyazima (1977) also makes the point that differences between the dialect divisions that people perceive and those that actually exist, but both are valid objects of study (see Preston and Long, 2002, p. 177-178). Onishi (2010) shows how 'geolinguistics' on Japanese varieties has moved away from the classification of broad dialects towards looking at the history of individual words.

In a computational model it may be more easy to specify and measure mutual intelligibility. For example, the number of words speakers know in common. If this is the case, however, one could use this measure directly to define linguistic communities dynamically and avoid coarse categories such as a 'language'.

3.7 Typology

The formalist approach to linguistics has developed a set of parameters to describe the features of a language. For example, basic word order, morphological case markings and so on. Differences in these formal properties could be used to differentiate languages. Indeed, Roeper (1999) sees bilingualism as existing wherever an individual has conflicting typological parameter settings. However, there are two problems. The first problem is that this approach uses a very different concept of bilingualism than that which is used in sociolinguistics or the one

defined in this thesis. Indeed, Roesper states that his sense of ‘Theoretical Bilingualism’ is “*orthogonal* to the obvious social dimensions of bilingualism which understandably have given predominant stature to the sociolinguistic perspective on bilingualism” (Roesper, 1999, p. 169, my emphasis). When it becomes clear that social contexts influence the choice of mini-grammar that speakers use, Roesper still maintains that social aspects of language are not part of linguistics:

“Why do languages have pockets of [Theoretical Bilingualism]? This would seem to be highly inefficient from a formal point of view. The answer, as we hinted above, may lie *outside of formal linguistics*. What makes a social register distinctive? What conveys to people the sense that a different level of communication is involved if, among bilingual speakers, one or the other language is chosen? These are deep questions which go *beyond linguistics* and my realm of expertise.”

(Roesper, 1999, p. 183, emphasis mine)

In another section, Roesper suggests that social registers are really also a part of UG: “Formal or Informal Speech Registers are recognizable as a choice of a different application of principles within UG” (Roesper, 1999, p. 183). Roesper then appears to offload part of the explanation of the irregularities of child speech onto this social register, usefully putting them beyond the scope of formal linguistics (Roesper, 1999, p. 183). Furthermore, Roesper’s approach seems to overestimate bilingualism. Individuals in a community can be in agreement that two conflicting grammatical options are perfectly grammatical without the community being considered ‘bilingual’ (e.g. Sorace, 2000). Indeed, Sorace sees knowledge about this kind of variation as a core part of linguistic competence.

The second problem relates to the psychological reality of categorical grammaticality - the basis for the formal parameters. Bard et al. (1996) show that grammaticality judgements are in fact more graded than previously assumed. Furthermore, Spruit (2006) show that grammaticality judgements change gradually over geographic areas rather than form more solid boundaries. If this is the case, then even formalist approaches cannot draw categorical boundaries between languages. On this basis, representing whole languages as a single unit in a model does not seem valid.

Finally, there is the reductionist argument. If we can specify the distances between languages using lower-level measurements, why not just use those in a model instead of the more abstract concept of ‘language’?

3.8 Descent

Just as biological species may be identified by their evolutionary descent, languages may be identified by their diachronic descent. For instance, English is

‘closer’ to Dutch than Norwegian because English and Dutch both descended from the West Germanic branch while Norwegian is part of the North Germanic branch. However, Hurford and Dediu (2009) argue that these kinds of taxonomies in linguistics conflate the descent of people and the descent of linguistic variation. While some studies have shown a correlation between genetic, geographic and linguistic descent (Novembre et al., 2008), in theory the two paths of descent may be totally independent⁶. For example, genetic descendants of African slaves in Jamaica speak a variety of English that is linguistically descended from Europe. If these are two separate mechanisms in the real world, then they should be separate in the model. That is, representing individuals independently from their linguistic system is important for a consideration of linguistic diversity, and a first step towards a valid model of bilingualism. Indeed, as Croft notes, “The replication of tokens of linguistic entities ... is mediated by speakers: linguistic entities cannot reproduce by themselves” (Croft, in press, section 1.2).

In addition to this, it is unclear how clean the descent of linguistic features really is. Mufwene (2005) argues that the descent of linguistic systems does not happen at the level of languages:

“Although some ethnographic considerations suggest that selection also applies at the level of languages, when speakers target primarily features of a particular language over those of others, what we know about language mixing and the development of creoles suggests otherwise. Languages are selected indirectly through the fact that their features (sounds, words, combinatoric rules, and particular ways of packaging meanings) wind up constituting the majority of those selected from the combined feature pool of the language varieties in contact. Although clearly favored, the indirectly selected language (variety) also bears the influence of (some of) the disfavored varieties and is therefore modified into a new variety.”

(Mufwene, 2005, p. 33)

This theory makes an empirically testable hypothesis: that sub-features of languages will exhibit different histories of descent. The next sub-section demonstrates this quantitatively, suggesting that descent fails to work as a criterion for identifying languages.

3.8.1 The descent of English words

Mufwene’s suggestion that linguistic systems combine in a piecemeal fashion can be quantified. For example, English is considered a Germanic language, but aspects (e.g. the lexicon) are heavily influenced by the other languages its speakers came into contact with and borrowed from (French, Welsh, Spanish etc.). Although the ‘last common ancestor’ of two languages such as English and Welsh

⁶(Hill, 1978, p. 3) also discusses the difficulty of defining discrete geographic ‘areas’.

might be Proto-Indo-European by some estimations, it is clear that they have been influencing each other much more recently. I attempt to quantify this by analysing an etymology dictionary of English.

The Online Etymology Dictionary (<http://www.etymonline.com>) is a digital collection of the roots of English words. The dictionary contains entries such as the following:

pace(1) “a step” late 13c., from O.Fr. *pas*, from L. *passus* “a step,” lit. pp. of *pandere* “to stretch (the leg), spread out,” from PIE **pat-no-*

Harper (2012a)

These entries were automatically parsed to give a descent of each word. For example, the entry for “pace” would result in a link from modern English to Old French to Latin to Proto-Indo-European. Parsing around 6,000 of these entries resulted in a graph with languages as nodes and borrowing links as edges. The resulting structure is more graph-like than tree-like, with many possible paths through language nodes to English (see figure 3.2, scale-free test $KS = 2.6904$, $p < 0.00001$)⁷. This demonstrates that ‘English’ has a complex descent, with lots of horizontal transfer. It also means that the same original word can have multiple routes into English, such as the Old Norse word ‘skrækja’ which is the origin for both ‘screech’ and ‘shriek’ in modern English (Harper, 2012b)⁸.

However, other studies that test the influence of horizontal transmission have concluded with the opposite response. Bower et al. (2011) show that lateral transmission does not account for a significant proportion of variance in the descent of basic vocabulary items. Simon Greenhill (private communication) points out that the level of analysis affects the answer to this question. Core vocabulary is likely to change less than the wider set of words considered in the etymology study above. Of course, as Hurford and Dediu (2009) suggest, many sources only contribute a few words to English, and may be ignored. However, it still leaves two branches that make significant contributions: Germanic and Romance.

Maybe we can use this graph of connections to demarcate a higher-order grouping from the bottom up. Graphs can be divided into statistically relevant clusters using network modularity (see Newman, 2006). An optimal modularity will split the graph to maximise the average degree (number of connections) within clusters. The network modularity categorisation for this graph splits the graph into three sections which align closely with Germanic, Romance and ‘other’ languages

⁷Nodes: 49, Edges: 182, Transitivity: 0.176, Shortest Path: 1.32, Weighted Shortest Path: 2.15 based on number of words borrowed, Average path length between any two nodes = 1.87, Assortativity: -0.375), Average Erdős number for modern English = 1.25. That is, language nodes are an average of 1.25 steps away from English. Average clustering coefficient = 0.244. Network diameter = 5 (largest path distance between two nodes).

⁸This example suggested by Jim Hurford.

(precision = 0.9, recall = 0.86, F = 0.88). That is, high-level linguistic categories are detectable just from low-level linguistic features.

There are two problems with categorisation through network modularity, however. Firstly, for tractability the optimal partition must be estimated (Newman, 2004) and so the clusters are statistical, not categorical. In this case, it seems like an unhelpful abstraction to use the categories when there is finer-grained information available. Secondly, the number of modules in which to divide the network is usually specified as a parameter. That is, the algorithm could find four, five or six clusters to specify a more fine-grained clustering. Although the categorisation in the graph fits with a language-family level analysis, lower-level categorisations may be just as valid.

Furthermore, the eliminativist argument works here too: If high-level categories can be defined through low-level linguistic features, then it is probably best to work with the low-level features. However, even in this analysis, the language nodes are abstractions - they represent waypoints in time and represent cultural communities as well as linguistic ones. All in all, it seems like an invalid abstraction to see whole languages as having tree-like descents. If this is the case, then defining languages through descent seems intractable. Certainly, if one were to model the diffusion of language on the basis of this analysis, encoding whole languages would seem invalid and unhelpful.

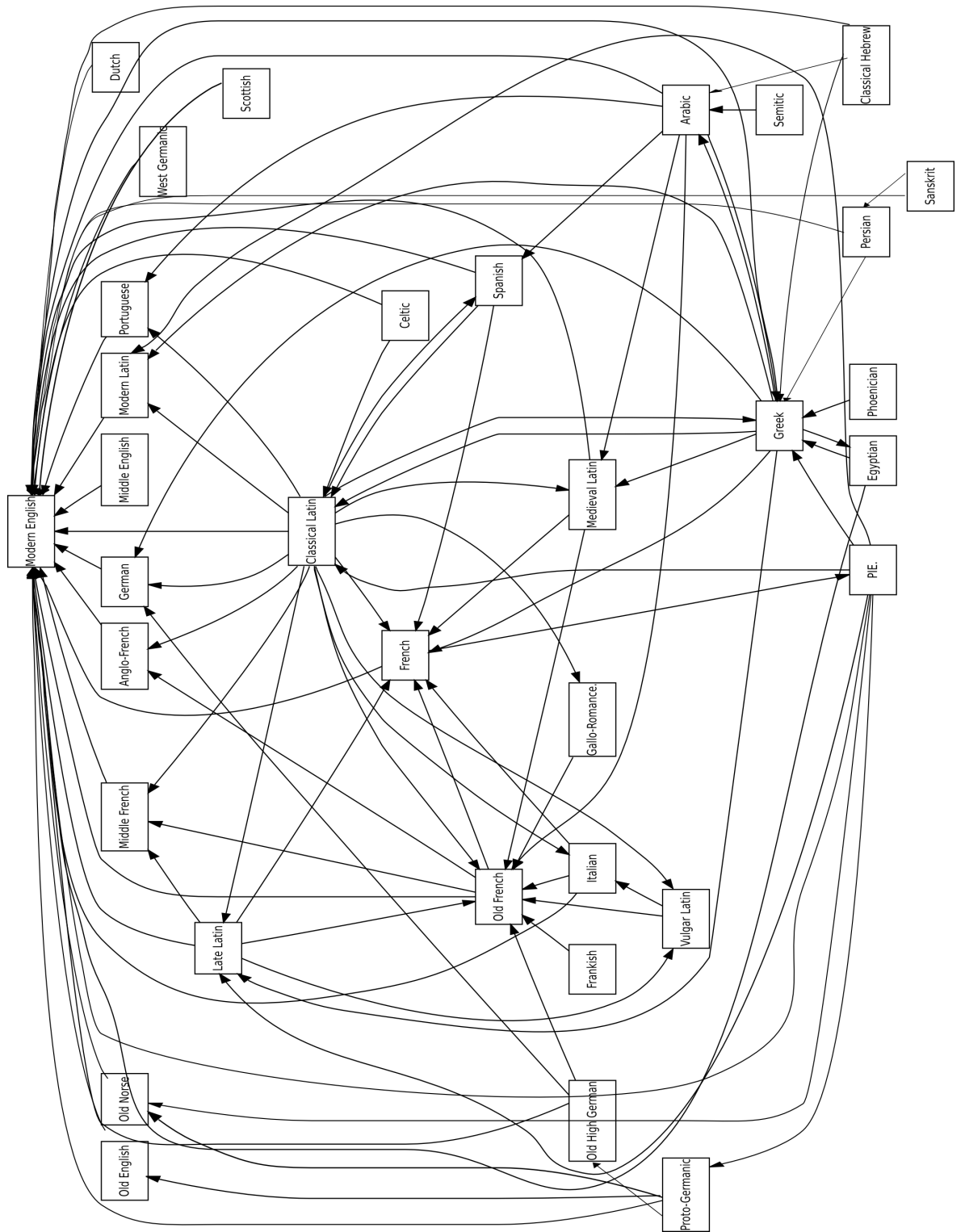


Figure 3.2: A graph of the descent of English words through other languages. Arrows point from source to borrower.

3.9 Statistical learning

Infants can use the statistical properties of their linguistic input to differentiate languages. The earliest differentiations are made according to rhythmic features, for instance syllable timing, stress timing and mora (e.g. Pelucchi et al., 2009 see Cutler, 2005 for a review). If infants can differentiate languages based on low-level features, perhaps languages are concrete, categorical units. However, the relationship between learning biases and languages is not so straightforward. Languages should adapt to the learning abilities of infants (Saffran et al., 2008), and infants adapt the cues they pay attention to based on the languages they are exposed to (Hay and Saffran, 2012; Cutler et al., 1989). Infants also exhibit similar statistical learning for non-linguistic stimuli (Saffran et al., 2007, 1999). This suggests that infants have low-level statistical learning abilities rather than high-level linguistic categorisations.

Some proponents of top-down models would argue that this is enough justification for building computational agents with internal representations of discrete languages (Griffiths and Kalish, 2007). In this case, I make two points. First, later chapters will show that different conclusions can come out of models where this is or is not the case. Secondly, the categories that would be delineated by statistical learning biases may not be useful for a study of bilingualism. Statistical biases might align with particular types of languages (e.g. based on rhythm types, see Goedemans and van der Hulst, 2011), and these might follow evolutionary laws that are amenable to study in such an abstract model. However, it is unlikely that these language types would consistently line up with the distributions of linguistic variation that were socially relevant. That is, infants can differentiate two distributions of linguistic variation in a given context, but this doesn't mean that these distributions are relevant at the population level in all situations across all time. Without a contrast in social variables, there may be nothing on which to condition the variation. Indeed, as the definition of bilingualism in this thesis suggests, without a salient difference at the social level, it seems meaningless or arbitrary to talk about many languages. Therefore, the statistical learning abilities of children suggest that models should not pre-specify what the relevant units of learning should be. Rather, computational agents should apply general statistical learning mechanisms to their input which identify relevant semantic features (e.g. see special issue of the *Journal of Child Language*, MacWhinney, 2010). The bottom up model presented in chapter 7 follows this suggestion.

3.10 Physiology

Maybe the previous measures are too high-level and we can get a tractable, qualitative measure of difference between language by looking at the hardware of language: the brain. Roux and Trémoulet (2002) report dissociations between languages using cortical stimulation of patients undergoing brain surgery. In or-

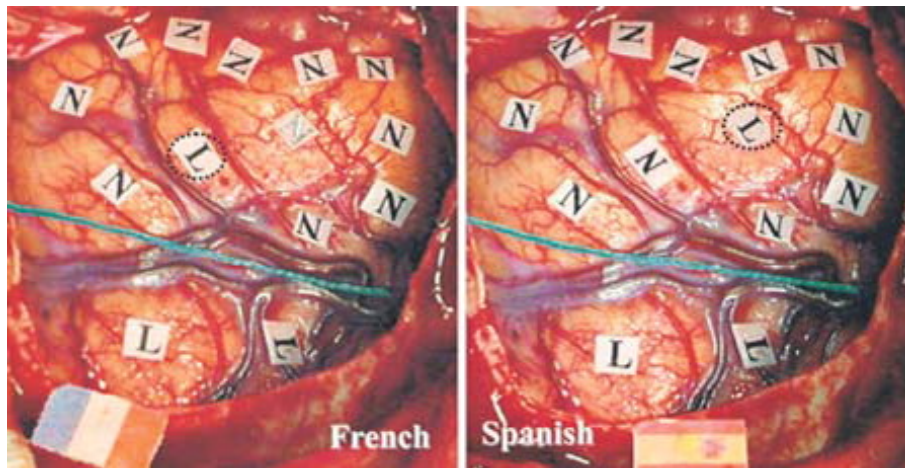


Figure 3.3: The exposed cortex of a French-Spanish bilingual being tested for speech arrest in French (left) and Spanish (right) from Roux and Trémoulet (2002). Areas labelled ‘N’ show normal speech activity. Areas marked ‘L’ are the locations of stimulation when speech arrest was observed. The dotted circles show two areas where speech arrest was observed for one language but not the other.

der to avoid damaging language-specific parts of the brain, surgeons first ‘map’ language-sensitive areas of the brain. This is done by exposing the brain while the patient is still awake and getting them to do a reading task while the surgeons stimulate parts of the cortex with small electrical currents. Stimulating some parts causes speech arrest - either loss of control of vocalising muscles or temporary anomia. Roux and Trémoulet report that stimulating areas of a bilingual’s brain can cause speech arrest (an inability to recall words) for words belonging to one language but not the other (see figure 3.3).

This dissociation between physical processing areas suggests a qualitative difference between languages. However, Roux and Trémoulet also report significant individual differences (see figure 3.4), including many cases of brain areas that caused speech arrest in more than one language. Pinning down how cortical areas come to differentiate between languages may be difficult because the relationship between cortical areas and language learning is poorly understood. Tests are yet to be done to see whether the dissociations between languages in the brain can also be reduced to lower-level differences. This may simply be a second-order problem.

Even if the brain is able to dissociate languages categorically, given that the other measures in this chapter fail to do so, we must ask on what basis the brain is able to accomplish this task. It is likely that non-linguistic factors are used such as context, speaker identity, cultural background and so on. If this is the case, then models should also use these features.

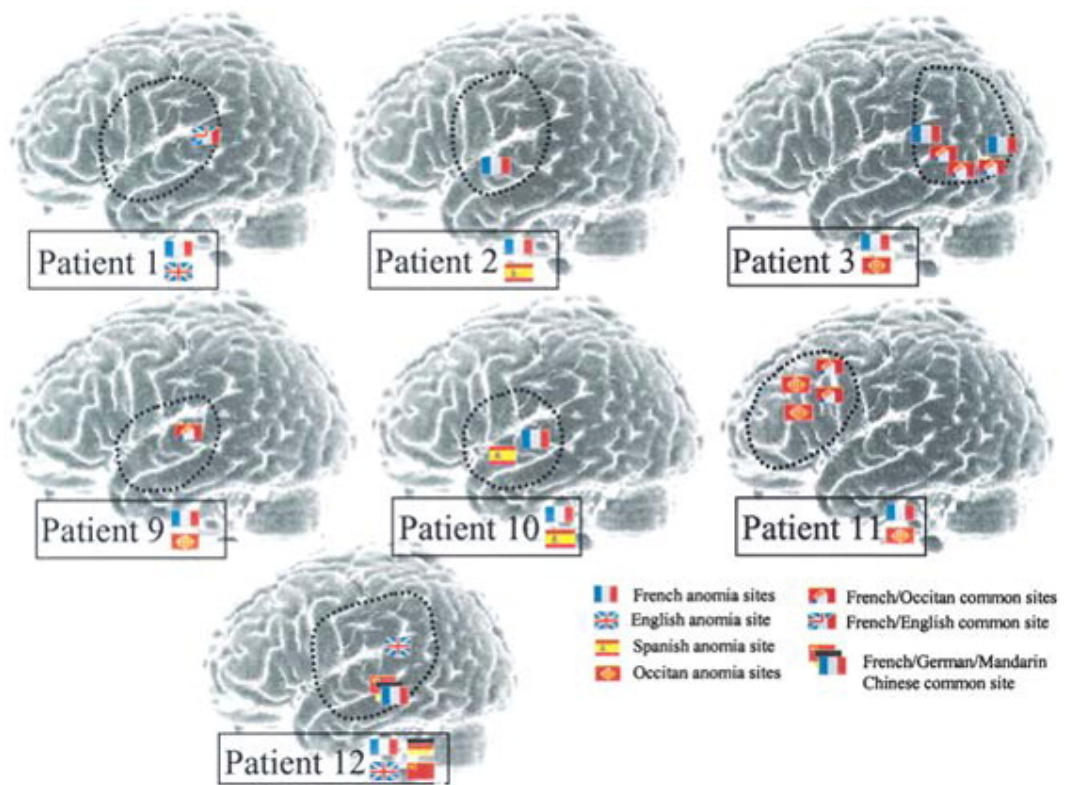


Figure 3.4: Sites of dissociations between languages for 7 patients from Roux and Trémoulet (2002). Dotted lines indicate the area of the brain exposed during surgery.

3.11 Function

The final measure considered here is functionality. The linguistic systems of individuals in isolated communities may have adapted to different cognitive niches, and so one might be able to differentiate languages on this basis. For example, Lupyan and Dale (2010) demonstrate a correlation between population size and the morphological complexity of their language. They explain this by hypothesising that larger communities are more likely to have more second language, adult speakers. Since adults find morphology more difficult to learn than children (Clahsen et al., 2010), populations with a greater number of adults will have a bias against morphological marking. Therefore, English - a global language with perhaps more second language speakers than first language speakers - is less morphologically complex than the language of a smaller, more isolated community such as Welsh. That is, the languages have adapted to different ‘linguistic niches’ (see Lupyan and Dale, 2010, see also Roberts and Winters, 2012 for a criticism of Lupyan and Dale’s hypothesis). One could appeal to the adaptive function of a linguistic system in order to try to categorise them into different languages.

However, this approach recognises that the differences between adaptive functions emerge from use and context, rather than being pre-specified. Furthermore, it recognises the impact that societies have on the language they use, suggesting that languages are better seen as population-level phenomena, not individual-level phenomena. Again, this does not support the use of an abstract concept of language in a model of language evolution.

3.12 Summary

The sections above have demonstrated that different ways of attempting to differentiate ‘languages’ are problematic. For instance, they may be inconsistent, be based on factors other than linguistic ones or just refer to low-level categories. The next sections consider the implications of this for models of language evolution. First, the argument of eliminating the concept of monolithic, static, discrete languages is considered.

3.13 Eliminating the concept of discrete languages

As we have seen above, different ways of measuring languages ignore the complexities of linguistic variation, for instance by ignoring how variation is distributed over speakers and concentrating on typological similarities. This section considers whether the concept of a language could be usefully eliminated from the field of linguistics. Craver (2007) argues that there are three reasons a concept may be invalid: underspecification, taxonomic errors and misidentification (see also Irvine, 2011). Concepts that meet these criteria may be usefully eliminated from the field (the concept should no longer be used to explain phenomena). Concepts

like aether in physics or the idea that the body is controlled by four ‘humours’ in medicine have been eliminated for this reason (see Whittaker, 1910; Arikha, 2008). The next three sections show that identifying ‘languages’ also leads to underspecification, taxonomic errors and misidentification.

3.13.1 Underspecification

A concept underspecifies the phenomena it tries to describe if it does not capture all its relevant aspects. The previous sections demonstrated that, while a ‘language’ may be treated as a linguistic concept, there are many considerations beyond purely linguistic ones that go into specifying languages. This suggests that the concept of a ‘language’ underspecifies the complexities of how linguistic systems vary between and within individuals.

3.13.2 Taxonomic errors

A language may be identified by linguists as a coherent linguistic unit, but in fact, as the sections above have shown, some divisions between languages are based on cultural or political differences. Also, as shown in the section on language use, speakers may treat varieties that others regard as separate languages as a single ‘medium’. Furthermore, it is possible that linguists group variation that is functionally differentiated, such as different ‘dialects’ under a single ‘language’. Also, the concept of a ‘language’ is often affected by the precise question of the researcher, as discussed in section 3.3. As Sober puts it “there is no linguistic justification for treating Danish and Swedish as different languages but Mandarin and Cantonese Chinese as different dialects of the same language” (Sober, 1980, p. 398).

The question of how categories like languages extend through time is also unclear. Clearly, there is a difference between the linguistic system currently spoken in London, and the language spoken there 3,000 years ago. Some of these changes might have been fairly abrupt due to invasion or migration, but without reference to social historical events, drawing lines between Old English, Middle English and Modern English is a fairly arbitrary exercise.

Conflicting with this descent-orientated categorisation are speaker-orientated concepts such as a ‘medium’ which is a group of languages treated by individuals as a single system (Gafaranga, 2008) and creoles which are a combination of two languages in contact. There are geographic concepts such ‘sprachbund’ which is a geographic area with closely related linguistic structures (see Thomason, 2000; Enfield, 2005), or ‘sprehbund’ which is a geographic area with closely related systems of use (see Sorensen, 1967, p. 677). Additionally, as section 3.5 showed, pre-existing definitions of language do not necessarily line up with the way speakers use linguistic variation. It seems that categorising linguistic variation into ‘languages’ often lumps together diverse linguistic systems, or divides

linguistic systems that are very similar. These are criteria for a concept making a ‘taxonomic error’.

3.13.3 Misidentification

According to Craver (2007), the elimination of a concept should only be considered seriously if it misidentifies its target phenomenon. If a concept is misidentified, then the phenomenon it refers to does not exist and using the concept is of little practical or explanatory use. It is possible that the concept of discrete, static languages is not useful for an evolutionary approach to bilingualism.

However, the concept of a language is clearly entrenched in linguistics, and it is clear that a lot of progress has been made using the concept. Eliminating entrenched concepts, especially ones where problems are widely acknowledged, might not be productive for the field (Brigandt, 2003; Ereshefsky, 2010). A project that seriously considered eliminating the concept of a language from linguistics is well outside the scope of this thesis.

Yet in terms of an evolutionary approach to bilingualism (how the distribution of linguistic variation changes over time), I do argue that there is no concrete analogue of a ‘language’ in the real world. This has been demonstrated by showing that there is no objective way of defining languages. This raises the question of whether the concept of a ‘language’ is useful for explaining processes of language evolution. In fact, chapter 8 will show that studies of bilingual language acquisition have started to describe their work without reference to high-level linguistic categories. Researchers have been focussing on how exposure to low-level linguistic variation affects language learning. If this trend continues, then perhaps the concept of a ‘language’ will be eliminated without explicit attempts to do so.

3.14 Requirements of a model of bilingual cultural evolution

In this chapter I have shown that many methods of measuring a categorical difference between languages are either invalid or rely on and emerge from lower-level distinctions. However, many measures of ‘bilingualism’ rely on these discrete, high-level categories. In this case, it is unclear what these measures of ‘bilingualism’ are actually measuring. So, even if the formal eliminativist argument does not hold, at the very least it is likely that the concept of a ‘language’ is very sensitive to the precise context of study or the research question of the researcher. This means that, if discrete languages are not valid abstract concepts, then they should not be encoded into models that explore bilingual cultural evolution. A ‘language’ in a model of language evolution should be a dynamic, context-dependent, emergent property of lower-level features. It follows that ‘bilingualism’ must also be

an emergent property that must be measured from the bottom up.

If there's no psychological reality to the concept of a language as a population-level phenomenon (as suggested by this chapter and by e.g. Chomsky, 1980), then there is an argument for not representing this unit in the minds of individuals in a model of cultural evolution that considers bilingualism. I argue that the concept of bilingualism only makes sense in a cultural evolution framework as an emergent, population-level phenomenon which can be measured as the amount of linguistic optionality that is conditioned on social variables. This view, together with the arguments from the current chapter, suggest three requirements for a model that studies bilingualism in a cultural evolution framework.

Firstly, that the abstract units of the model be low-level linguistic features, not whole languages. For instance, it is not valid to simply represent an individual as knowing a language or not as a binary feature. It is valid, however, to represent an individual as knowing a linguistic feature or features such as a phoneme inventory, lexical items, or syntactic features such as headedness or basic word order, assuming that these can be linked to specific cognitive processes in the individual.

Secondly, models should represent individuals explicitly. Some models confound the concepts of an individual and their linguistic system (see the next chapter). While this does not matter for a model of monolingual speakers, a difference emerges when individuals can adopt many varieties. I argue that the concept of bilingualism relies on how language is used between individuals. Therefore, models should represent a concrete number of individuals and tie specific linguistic utterances to specific individuals.

Finally, the social structure in these models should be complex and dynamic. By complex, I mean that individuals do not necessarily have an equal probability of hearing an utterance from any other individual, within and between generations. The sociolinguistic literature suggests that imbalances in and changes to the social structures are part of the driving force for linguistic change (e.g. Kaufman and Thomason, 1988; Mufwene, 2001; Winford, 2003). Wichmann and Holman (2009) also argue that complex social structures should be a key feature in a model of language change. The most extreme example is two populations that are entirely isolated. Other cases are possible too, such as a population having two sub-populations that are more likely to communicate within their sub-population than between it. Imbalances in the probability of communication are also important. For example, consider a situation with a dominant majority population and a minority population. A real world example might be an immigrant community who have retained their native language as well as adopting the host community's language. Individuals in the minority population might communicate equally often with individuals from either community, but individuals from the majority community might be much less likely to communicate with an individual from the minority community than the majority community.

Social structures should also be dynamic. Linguistic diversity is driven by contact between groups (e.g. Nettle, 1999a), which involves the social structures changing. For example, two communities that come into contact move from isolation to being more or less integrated. Communities may also split and diverge. Models that consider linguistic diversity over evolutionary time (many generations) should consider the possibility that social structures can change. This is not a new argument. Modelling social structures as static and endogenous (where descent happens only within groups) has been criticised in other fields such as human ecology (see Lee, 1972). Jane H. Hill also argues as far back as 1978 that dynamic social structures are a key aspect of understanding areal phenomena, multilingualism, pidgins, diglossia and critical period effects (Hill, 1978). Hill suggests individuals can adapt - in their language, behaviour and genetically - to structures beyond the local group. For instance, the Yanomama, though they live in an area that appears to be conducive to linguistic homogeneity, speak several languages and have a commonly spoken 'high' variety that is used for communicating between local groups (see Hill, 1978, p. 11-12). Despite this, what Hill would identify as the 'dialect tribe' model have become the norm in computational models of the cultural evolution of language (see chapter 5). While models must make simplifying assumptions to understand general processes, I argue that dynamic populations are a central feature for the study of bilingual cultural evolution.

If these three requirements are met, then the issue of discrete languages can be circumvented. The next issue is how to measure bilingualism in a model without discrete languages. Following the definition of bilingualism developed in this thesis, bilingualism can be measured as the amount of linguistic variation that is conditioned on social factors. This only requires low-level variation and contrast in social variables, which the three requirements above ensure. These requirements are implemented in a bottom up model presented in chapter 7.

3.15 Conclusion

This chapter argued that there is no consistent way to divide linguistic variation into discrete, higher-level categories like 'languages' on a purely linguistic basis. This might suggest that 'languages' do not have a psychological reality in the minds of learners, and so models of bilingualism should not encode discrete languages. Three requirements for valid models of bilingualism were suggested. These will be used to evaluate current models of bilingual cultural evolution, and also to motivate a bottom up model of bilingualism in chapter 7. However, if 'languages' are epiphenomena - categories that emerge from lower-level phenomena - then 'bilingualism' as the ability to speak more than one 'language' may also be an epiphenomena. The next chapter considers this issue and argues that bilingualism is still a valid object of study for evolutionary linguistics.

CHAPTER 4

INDIVIDUAL BIASES AND POPULATION LEVEL PHENOMENA

“I’d love it if I were bilingual, I’d love it if I were multilingual. If I had many different languages, I’d be a richer person.”

Noam Chomsky (in Baron Cohen, 2004)

4.1 Introduction

The previous chapter argued that ‘languages’ are not monolithic, static, discrete entities, but can be seen as ‘epiphenomena’ that emerge from use, culture, politics, identity and so on. This suggests that representing discrete languages in models of bilingual cultural evolution may not be valid. However, by arguing for the elimination of this view of languages, a problem arises: if bilingualism is knowing more than two ‘languages’, then bilingualism may also be an epiphenomenon, and so subject to the same eliminativist argument. It has been argued that this is the case, and so bilingualism is not a productive object of study for evolutionary linguistics (see Chomsky, 1980, 2000; Sober, 1980).

I oppose this position in two ways. First, it is possible to measure bilingualism without using a discrete notion of languages. This is based on the amount of linguistic variation that is conditioned on social factors, as described in the first chapter. This measure will be implemented in a model in chapter 7.

In this chapter, I oppose the eliminativist case for bilingualism by using arguments from biology about the relevance of the concept of a ‘species’ (Sober, 1980). The parallel that I’d like to draw between biological and cultural evolution for the purposes of this thesis is a historical one: The *kind* of problem that biology faces when trying to study speciation is the *kind* of problem that linguists face when studying linguistic diversity. That is, how to think about low-level features that are in competition but that are grouped into individuals, who themselves are part of larger groups that are defined by, but also emerge from the low-level features.

I argue, in parallel with Sober (1980) that just because ‘languages’ do not have a psychological reality, it does not mean that bilingualism cannot be a valid object of study in its own right. If bilingualism at the population can be shown to

exhibit law-like behaviour, then it can be studied in its own right, and may provide insights into questions about language evolution. The aim of the bottom-up model in chapter 7 will be to demonstrate this point.

The first section in this chapter discusses the idea of conceptualising languages as species, and parallels between problems in biology and linguistics. The second section discusses the argument that bilingualism is an epiphenomena, and therefore not relevant to the study of language evolution. The final section discusses the implications for abstract models of cultural evolution.

4.2 Languages as species

One of the conceptual problems in this thesis is how to relate concepts at the individual level to concepts at the population level. This has also been a problem for biology. The debate over whether the concepts and approaches of evolution can be applied to language is as old as the theory of evolution. Darwin realised that languages might be subject to evolutionary pressures (Darwin and Ghiselin, 1874) and a contemporary, Schleicher, was quick to start pointing out the similarities and differences (Schleicher, 1863).

Mufwene (2001) promotes a view of language as an analogue of species (rather than an organism e.g. Christiansen and Chater, 2008) based on three points: Languages are collections of idiolects, while idiolects are more like organisms; Languages die in protracted ways, unlike organisms; Languages are acquired in a piecemeal fashion. Although animals don't fit this view, other organisms such as viruses do change their genetic makeup over their lifetime.

The problem of identifying languages has many parallels with identifying a 'species' in biology (see Ereshefsky, 1998). The most relevant definition may be based on the ability to mate (in linguistics analogous to mutual intelligibility), genetic similarity (typological similarity), phenotypic similarity (phonetic similarity), evolutionary history (cultural history), behaviour (language use) or ecological niche (cognitive niche). Some of these are discussed in the previous chapter. These concepts are problematic in biology as well as linguistics. For example, a 'ring species' (Mayr, 1942) is a group of sub-species where each sub-species can mate with their neighbouring type, but not necessarily with every member of the group (e.g. *Larus* gulls, Haffer, 1982, although see Liebers et al., 2004). Drawing a species boundaries in this situation based on mating is difficult, and there are parallel situations in linguistics (Sober, 1980; Nettle, 1999a and see section 3.6 on page 34 on dialects in Japan). For instance, Radcliffe-Brown (1930) finds that an indigenous community in Australia might use many dialects, or a single dialect might span many communities, concluding that "it is difficult to say exactly where one language ends and another begins" (Radcliffe-Brown, 1930, p. 37).

However, there are some problems raised by bilingualism that potentially confuse the language-as-species analogy. First, while many conceptions of separate species rely on the inability to mate, as Diebold (1961) points out in his essay on ‘incipient bilingualism’ that “biological change is almost totally divergent. Linguistic and cultural change, on the other hand, is rarely if ever free from convergence. The analogue of biological speciation in linguistic change is language and dialect formation; but mutual unintelligibility, unlike mutual infertility, is no block to systemic convergence.” That is, learners can learn any combination of languages, regardless of how divergent the languages have become. Either this means that there has been no divergence in languages - that there is only one language species - or that a different process is involved. The situation becomes more complicated if one considers bilinguals. If languages are species, what is the analogue of a bilingual individual? Two separate organisms? An organism with two sets of genes? Do two language ‘organisms’ inhabit one brain-ecology? This seems to go against studies of bilingualism and language mixing that show that an individual has a single systematic linguistic system (e.g. Gafaranga, 2007).

The nature of this problem has two origins. The first is that a cultural, linguistic system is being pressed into an analogy with biological evolution. While it is perfectly reasonable to assume that linguistic variants are subject to evolutionary processes, this does not imply that there will be analogues at all levels of analysis.

However, there is another kind of problem that is faced by linguistics which is not such a concern for evolutionary biologists. This is the argument of this chapter: I suggest that part of the confusion of how to represent linguistic structures in an evolutionary framework is actually based on the invalid concept of a language, or rather as an assumption that a bilingual is qualitatively different from a monolingual (or indeed that the monolingual is the ‘default’). In later chapters, I will argue that rather than needing to count the number of systems, we can just assume that each individual has a learning system that conditions linguistic variance on semantic variables. All individuals exhibit variation, but if socio-cultural variables are important for a particular group of individuals, then the phenomenon we recognise as bilingualism should emerge.

This approach does not rely on specific analogues with evolutionary biology, but does use the trick of disregarding high-level categorisations that are imposed by researchers and instead seeing how categories emerge from low-level variation, interactions with the environment and social structure.

4.3 Epiphenomena

The previous chapter argued that ‘languages’ are epiphenomena. That is, a property that is caused by certain configurations of lower-level features, but not something that had a consistent ‘psychological reality’ in the minds of speakers.

By extension, ‘bilingualism’ as conceptualised as knowledge of more than one language, is also an epiphenomena. Chomsky suggests that this means that bilingualism is not a productive object of study for evolutionary linguistics:

“This branch of the study of language [Generativism] is indeed marked by an absence of any role for community and culture ... There is nothing of any significance known, at least to me, about community and culture that relates to these questions about the nature of a certain biological system. ... Bilingualism is normal to the species in the trivial sense that the world is so complex that strict monolingualism is almost unimaginable. In that sense it is natural to the species but I don’t see anything deep about that.”

Chomsky (2000)

Chomsky suggests that linguists should focus on language in the individual rather than language in a society. Sober (1980) relates the line-drawing problem in linguistics to the reasons why Chomsky (1980) opposes the study of languages as social entities. First, it is difficult to draw lines between dialects synchronically or diachronically, making it difficult to state criteria for distinguishing objects of study. Chomsky also sees languages as properties of groups, as ‘epiphenomena’ arising from interactions of individuals. That is, they are effects, not causes. Sober characterises Chomsky’s argument as “so even if line-drawing problems could be solved, there still would not be a reason for thinking that socially shared languages are needed to explain anything” (Sober, 1980, p.397).

With regards to the second argument that population-level languages are just epiphenomena, Sober notes that this is a reductionist argument. That is, the linguistic representations of individuals are ‘epiphenomena’ of lower-level brain processes, yet this does not mean that studying individual psychology is invalid. By extension, then, the fact that population-level phenomena in linguistics are ‘epiphenomena’ of the interaction between individuals does not mean that population-level phenomena cannot be studied in their own right.

With regards to the first argument about the line-drawing problem, Sober notes that the same problems exist in defining biological species, yet a ‘species’ is a generally accepted concept in biology. This is possible because of ‘population thinking’ (Mayr, 1963). Populations obey evolutionary laws, so there might be no need to define them in terms of their constituents. Sober suggests that it is possible to see languages as social entities and linguistics as the study of what individuals in speech communities have in common. It is then possible to prove that population-level phenomena are real by describing the laws that they obey:

“To show that properties of social wholes are not simply artefacts of the properties of their parts, one must show how such social properties are connected to each other in a law-like way. ... Given such

interconnections, particular events can be explained in terms of social properties and their laws; this sort of explanation entails that social properties are not mere epiphenomena, since the explanation will attribute causal efficacy to populations' possessing the properties they do. ... only the elaboration of empirical theories can fully vindicate the idea of social reality"

(Sober, 1980, p. 404)

The literature on bilingualism might provide some evidence. For example, the learning of a second language is affected by how it relates to a learner's first language (see De Groot, 2010). Also, exposure to multiple 'languages' (socially defined clusters of linguistic variation) affect individual cognition. For instance, executive control profiles (Treccani et al., 2009; Hernández et al., 2010), word learning heuristics (Houston-Price et al., 2010; Byers-Heinlein and Werker, 2009a; Healey and Skarabela, 2009; Merriman and Kutlesic, 1993) or meta-linguistic knowledge (Bialystok, 1991; Serratrice et al., 2009). 'Bilingualism', then, seems to be a useful concept for describing effects on individual cognition. This co-evolution is discussed further in chapter 8.

However, as the previous chapter argued, defining bilingualism in terms of knowledge of the number of discrete languages a person 'knows' is problematic. The issue of encoding languages in models of cultural evolution is now discussed.

4.4 Abstractions in models

Abstract models of cultural evolution have also been looking at the interaction between individual biases and population-level phenomena. For instance, under certain assumptions, the distribution of linguistic types in a population will converge to the prior biases of its individuals (Griffiths and Kalish, 2007). This is a good demonstration that linguistic distributions at the population level are epiphenomena, but not necessarily that they have causal effects. However, some recent models allow the biases of individuals to adapt to the population-level phenomena, in a similar way to the evidence from studies of bilingualism might suggest (Smith and Thompson, 2012; Thompson, 2012; Thompson et al., 2012). The bottom-up model presented in chapter 7 also allows the linguistic representations of individuals to be influenced by population-level phenomena.

Some models of cultural evolution are designed to oppose the claim that there are transparent links between population-level phenomena and individual biases (Kirby et al., 2007; Thompson et al., 2012). The models demonstrate that observing strong linguistic universals in the world does not necessarily mean that individuals have strong language-specific biases. Rather, cultural transmission can amplify weak, general biases. However, by setting the models up to oppose strong nativist arguments, they have assumed that there is a psychological reality to high-level linguistic categories. That is, the model has individuals who

share internal states that are monolithic, discrete and fixed across time and the population-level phenomena are direct measures of these internal states. Chomsky might argue against the point about the relationship between universals by claiming that linguistic systems of populations do not have a psychological reality (Chomsky, 1980) and so social forces are uninformative for the study of the internal representation of the individual (Chomsky, 2000).

The issue of the transparency of the relationship between individuals' internal representations and the population-level phenomenon is brought into sharp focus when trying to model bilingualism. In the top-down models mentioned above (and discussed further in chapter 5), the population-level phenomenon is simply a count of the internal representations of individuals. Put another way, individuals have internal representations of the population-level phenomena. It is unclear whether this really does support the study of social phenomena. Furthermore, models that represent whole 'languages' as discrete, static entities assume that the line-drawing problem can be solved.

The previous chapter argued that 'languages' are complex, population-level phenomena that can only be identified through the way individuals use them in a particular context. In this case, individuals are unlikely to have evolved biases over discrete 'languages'. It may not even be possible to define a universal set of discrete 'languages' that align with learned internal representations. Therefore, for models that are designed for studying bilingualism in a cultural evolution framework, it might be invalid to encode individual representations as being essentially not distinct from the population-level phenomena. In response to this, a bottom up model is presented in chapter 7 where the individual representations and the population level phenomenon (bilingualism) are clearly separate.

However, this is a complex issue. Many models that make the assumptions above are not specifically designed to answer questions about bilingualism. Furthermore, the authors argue that the dynamics of the model only need to approximate the real world at some level for them to be insightful (e.g. Griffiths and Kalish, 2007, p.450-451). Similarly, Ke (2004) recognises that languages are complex entities, but argues that abstract languages may still be analysed at the population level:

“In linguistics, however, very often an abstract language system is taken as the object of analysis. This level of analysis disregards the distinction between idiolect and communal language, and neglects the heterogeneous nature of language at both levels. As a consequence, explanations for observed patterns based on this abstract level of analysis are often inadequate. However, this is a necessary step for linguists to identify interesting phenomena in the first place. At this abstract level of analysis, the self-organization framework can also be applied. ”

(Ke, 2004, p. i)

Nettle suggests that this argument also applies to linguistic diversity in dynamic social networks:

“It is clear ... that the formation of any particular ethnolinguistic group will be a complex interplay of many locally specific factors; formation of social bonds will depend upon precise topographical, military, epidemiological, demographic, and cultural situations, as well as more nebulous contingencies such as the rise and fall of local prestige and influence. However, I believe general explanations are appropriate for the global trends.”

(Nettle, 1998, p. 361)

This suggests that the level of abstraction of the model should be guided by the questions of the researcher, rather than aim to model reality perfectly. Many of the cultural evolution models see the linguistic units as representing low-level features such as word order, rather than whole languages. It is conceivable that individuals really do have cognitive biases over these units (e.g. Diamond, 1991, p. 143). The point of contention is whether the distributions in the world reflect the *strength* of these cognitive biases in a straightforward way (Kirby et al., 2007; Thompson et al., 2012). The difference between modelling these lower-level features and modelling bilingualism is that bilingualism is a purely social phenomenon. That is, you can't identify whether an individual is bilingual by studying them in isolation. Bilingualism can only be measured through the interactions of individuals. In contrast, the word order that a speaker used would be identifiable by their utterances alone. It is less clear, then, whether bilingualism at the population level relates straightforwardly to an individual cognitive bias. A model that assumed that there was a relevant cognitive bias (e.g. for the amount of variation to expect in your input, see Burkett and Griffiths, 2010, discussed in chapter 5) already makes a strong assumption about individual cognition.

Rather than attempt to construct a philosophical argument that addresses this issue, this thesis instead demonstrates that two approaches can lead researchers to different conclusions. This is done by comparing the results of two models. The first is a top down model, which assumes that learners have a cognitive bias over the distribution of variation to expect (chapter 6). The second model is a bottom up model that assumes there is a general statistical learning mechanism that is sensitive to social cues (chapter 7). The bottom up model also uses a measure of bilingualism that is based on how individuals use low-level variation. Using this kind of measure means that bilingualism is not based on the concept of discrete languages.

Both models will suggest slightly different resolutions to the bilingual paradox. The top down model suggests that the prevalence of bilingualism does not need to be underpinned by a strong bias to expect bilingualism. On the other hand, the bottom up model suggests that bilingualism emerges due to dynamic social

structures which support contrasts in social identity. Furthermore, it is possible to measure bilingualism in the bottom up model without reference to discrete internal representations of individuals. At the very least, then, it seems that exploring this problem from many perspectives is prudent.

4.5 Conclusion

This chapter argued that bilingualism can be studied as a phenomenon in its own right. Doing so would involve showing that the phenomenon of ‘bilingualism’ exhibited law-like behaviour. This can be seen in the effect of variation in the input on cognition (see above and chapter 8). It could also be demonstrated by constructing a model to study the dynamics of bilingualism. The next few chapters look at two types of model that do this. However, the first type, top down models, usually assume that individuals have internal representations of discrete ‘languages’, which have been shown to be problematic in this chapter the previous chapter. This motivates a bottom-up model that uses a concrete measure of bilingualism. Instead of being measured as a count of internal representations, bilingualism in the bottom up model can be measured by looking at the amount of low-level linguistic variation that is conditioned on social variables.

CHAPTER 5

TOP-DOWN APPROACHES TO ITERATED LEARNING

“Bir dil bir insan, iki dil iki insan”

One who speaks only one language is one person, but one who speaks two languages is two people.

Turkish proverb

5.1 Introduction

The previous chapters argued that bilingualism should not be thought about as the ability to speak multiple discrete, static languages. Instead, bilingualism was defined as a property of populations based on how linguistic variation is conditioned on social variables. The previous chapter suggested some requirements for models investigating bilingualism under the framework of the cultural evolution of language. The current chapter reviews some models of cultural evolution that are related to bilingualism in the light of these requirements. I will argue that the majority of models have some kind of assumption that monolingualism is the goal of the learners or the most normal, rational expectation for a learner to have. Even a cultural evolution model that is designed to study learning multiple languages from multiple teachers (Burkett and Griffiths, 2010) may not be entirely suitable for studying bilingual cultural evolution. A challenge for top down models is to continue to produce solid results for increasingly complex models. The final section discusses the literature on flexible learning mechanisms from studies of bilingualism that would be desirable to model.

Sections 2 and 3 review some mathematical and agent-based models of cultural evolution and point out biases against bilingualism. The bulk of this chapter discusses models of iterated learning in some amount of technical detail. The main focus will be Bayesian models of iterated learning. For readers who are not immersed in this literature, sections 5.4 and 5.5 provides a brief introduction to this kind of model, and a summary of the conclusions of these models relating to bilingualism. Section 6 reviews Burkett and Griffiths (2010)’s model which includes agents who learn multiple ‘languages’ from multiple teachers. These agents have expectations about the number of languages they will observe in their input. These expectations turn out to be important for the dynamics of the model. Sec-

tions 7 and 8 review some extensions of this model that allow these expectations to be learned from the input or evolve biologically. It will be demonstrated that these, too, tend to make assumptions that limit the possibility of bilingualism. After a brief summary in section 9, section 10 discusses the problems of representing bilingualism in top down models. It will be argued that current top down models need a more sophisticated conception of bilingualism. However, the challenge to top down models is not necessarily to reflect real systems more closely, but to continue to produce solid results for increasingly complex models.

5.2 Mathematical models

Abrams and Strogatz (2003) (extended in Stauffer et al., 2007) present a mathematical model of language death. It assumed that there is a population of agents and two possible languages. Agents decide to adopt a language based on the proportion of people who speak it and the cultural status of the language. Over many generations, one language tends to dominate. The model was fitted to real data of the decline of minority language speakers and, with some tuning of the parameters, fit reasonably well¹. Castelló et al. (2008) also extended the model to include different social structures². More complex structures extended the time the system took to reach uniformity. However, in populations with ‘community structure’³, where agents could adopt either language or both, linguistic diversity could persist indefinitely. Sub-communities in the population that are poorly connected to the rest of the population were resistant to being ‘invaded’ by the other language.

The problem in this model, for the purposes of this paper, is the way language is represented (also discussed in Fernando et al., 2010). Languages are monolithic entities - agents either know a language or they do not. In fact, in the original model, individuals could only know one language, so bilingualism was impossible. Furthermore, there is no scope for the perception of the division between languages to change based on social factors. As I have argued in chapter 3, languages are social constructs that are continuously re-constructed from generation to generation, rather than options that are fixed through time. While some

¹Although Fernando et al. (2010) point out it is a phenomenological model rather than a mechanistic model that can fit the data but not explain what the mechanism behind the behaviour is. Furthermore, Stauffer et al. (2007) demonstrates that an agent-based model of the same process provides qualitatively different results to the mathematical model. Reali and Griffiths (2010) also demonstrate that s-curves in the shifting of frequencies can be observed without assumptions about prestige, see section 5.9.1.

²This was an agent-based model where individuals could speak one of two languages or both, and the two languages were socially equivalent.

³This refers to the following condition: “Social networks are organized into communities with dense internal connections, giving rise to high values of the clustering coefficient. In addition, these networks have been observed to be assortative, i.e., highly connected vertices tend to connect to other highly connected vertices, and have broad degree distributions.” (Toivonen et al., 2006, p. 851)

linguistic features may be represented in a binary way (e.g. the basic order of subject and object), languages are much more complicated entities. That is, the measure of ‘bilingualism’ in this model is not equivalent to the measure suggested by this thesis (i.e. the amount of linguistic variation conditioned on social structures). Although much more complex models have been constructed which allow more realistic social aspects (e.g. Castelló et al., 2008; Fernando et al., 2010), they still maintain a discrete division between languages.

The second kind of problem that this model makes is a misunderstanding of the key aspects of language change. This model is typical of many physics models that look at the quiescence of a process in a closed system (in the thermodynamic sense). That is, the system is perturbed once and the resulting oscillations are studied before the system converges. However, one of the most obvious points from the bilingualism and sociolinguistics literature is that linguistic change is driven by changes in population or social structure or in cases of language contact (in physics terms, when energy is injected into the system). Indeed, Mufwene (2005) assumes as a first principle of trying to understand what causes linguistic change that “variation within a population is likely to remain stable unless something happens in its external ecology that disturbs the ‘balance of power’ between competing variants” (Mufwene, 2005, p. 23). The model is a weak representation of the process of linguistic change because it does not allow the interaction of social structure and language structure. Indeed, Abrams and Strogatz (2003) were forced to admit that “contrary to the models stark prediction, bilingual societies do, in fact, exist” (Abrams and Strogatz, 2003, p. 900).

Although not directly connected with the main literature in this chapter, and although it has very different aims, another mathematical model deserves some attention. Fernando et al. (2010) present a mathematical model of the dynamics of language death. They note that Abrams and Strogatz, 2003’s model does not suggest any mechanisms by which the decline of the minority language emerges. In order to rectify this, a much more complex model is constructed. Individuals speak either a low-status language (L), a high status language (H) or are bilingual (B). Individuals mate to produce children, though speakers of H cannot mate with speakers of L (although see Piller, 2000). Therefore, parents can be of four types: HH, LL, HB or LB. Children inherit their language state based on the frequency of languages they are exposed to, which is determined both by the state of their parents and the frequencies of languages in the community⁴. This process is iterated to produce generations of learners. The proportions of each speaker type is tracked over time. There are seven parameters that determine the dynamics of the model:

1. The initial proportions of L and H
2. The probability that a child of LL or LB parents will acquire H based on what they hear from the community

⁴Fernando et al. (2010) note that the second process is Lamarckian rather than Mendelian.

3. The probability that a child of HH or HB parents will acquire L based on what they hear from the community
4. The ratio of data received from parents and the wider community
5. The amplification of H heard due to exposure to public sources (e.g. Television).
6. The proportion of conversations heard by a child as a result of government intervention (e.g. education)
7. The rate at which children who learn only H can be taught L.

For each of these parameters, Fernando et al. suggest a plausible range and the data that would be required to estimate the parameters in the real world. For instance, variables 2 and 3 above could be estimated from data from De Houwer (2007). The baseline dynamics of the model demonstrate that L declines in frequency until it becomes extinct, as in Abrams and Strogatz (2003). However, by manipulating the parameters, Fernando et al. demonstrate that prestige (the third parameter above) actually has little impact on the maintenance of L. Better strategies for maintaining L include increasing the amount of exposure to the low-status language (variable 6 above) and teaching the low-status language to children who speak the high-status language (variable 7 above). These interventions increase the number of bilingual speakers (and couples where one parent is bilingual), rather than monolingual L speakers.

The model includes complex, dynamic social structures, but still only includes two monolithic, discrete, static languages. Also, the social structures are not independent from the linguistic identity of the speakers, so more parameters would be needed to model language contact situations. Furthermore, the individuals have no cultural identity separate from their linguistic identity, so one cannot measure the amount of linguistic variation that is conditioned on social variables. The strategies for maintaining L lead to a community where around a third of the population are bilingual and half are monolingual. For the kind of timescale the model aims at describing (500 years, see Fernando et al., 2010, p.66), it is unclear whether the contrast between languages and the prestige of H could be maintained. Certainly, this model has little to say about how divisions between languages occur in the first place. However, the scope and purpose of the model is clear and fit the research question well. The implications of this model on bilingual language policy are discussed in chapter 9.

5.3 Agent based approaches

Agent based models have been used to study how linguistic variation in a population evolves (e.g. Steels, 1996; Nowak and Krakauer, 1999; Briscoe, 2002; Steels and Belpaeme, 2005; Vogt, 2005; Gong et al., 2008; Puglisi et al., 2008; Loreto

et al., 2010; Baronchelli et al., 2010; Gong and Wang, 2010). A full literature review is outside the scope of this thesis. However, one paradigm is discussed here because of its assumptions about bilingualism. The minimal naming game (Gong et al., 2008; Puglisi et al., 2008; Loreto et al., 2010; Baronchelli et al., 2010) is a model where individuals try to converge on common labels for common categories for a continuous meaning space. The paradigm is often couched in terms for learning categorical boundaries for colours and their corresponding names. Individuals do this by playing guessing games in pairs. The agents are presented with a context of a sample of colours. One agent refers to one of the colours using a label and the other agent must guess which colour the first agent was referring to. Depending on whether the guess is correct or not, the memories and categorical boundaries of the agents is updated.

The interesting factor for this thesis is the assumptions that are made about this updating process. The minimal naming game is set up so that the goal of the system is uniformity between individuals. This goal is typically approached by trying to limit the amount of linguistic variation in the system. In Roberts (in press, included in section D.1 in appendix D), I argue that the updating algorithm implements a mutual exclusivity principle. For instance, if the guesser guesses the correct referent, both agents discard all other ‘competing’ labels for that referent’s colour category. The measurement of the progress of the system also usually involves a measure of increasing uniformity and decreasing or bounded linguistic variation (also used in other types of model, e.g. Castelló et al., 2008). However, decreasing the variation is not the only logical route to uniformity. Uniformity can also be achieved by distributing the variation evenly (i.e. if agents remember all labels they encounter then uniformity will also decrease over time). In Roberts (in press) I show that the mutual exclusivity features are not necessary for uniformity, and in complex social structures can actually inhibit uniformity. The social structures are an important factor, then, in the route the system takes to uniformity. Indeed, diversity can emerge when agents are spatially distributed (Steels and McIntyre, 1999), or when the social network adapts over time so that individuals who communicate successfully are more likely to interact, even with the standard minimal naming game algorithm (Lipowska and Lipowski, 2011).

A researcher interested in bilingualism, then, might be biased to start looking at complex social structures immediately. This can lead to further differences in the approach to the model. For example, the model looks at how the system reaches an efficient state, but there is more than one way to approach efficiency, too. In a fully connected social network, the communicative efficiency can be measured by the number of labels each individual needs to store. However, in a more structured network communicative where some agents don’t need to communicate effectively with some others, the communicative efficiency might need to be based on actual communication success (see Steels and McIntyre, 1999; Gong and Wang, 2010).

The choice of social structure also affects some more basic choices in the model such as how interactions are managed. For example, in a regular network (with an even number of agents), pairs of agents can interact simultaneously so there are discrete interaction phases. However, this is not necessarily possible in a (non-probabilistic) social structure where there may be imbalances in the number of neighbours that agents have (imagine a hub with three satellites: not everyone can talk to a neighbour at the same time).

In general, then, considering bilingualism can affect the features that are desirable in a model, including the assumptions about the learning biases of individuals, the parameters of the model and the measures of the progress of the model.

5.4 Introduction to Bayesian modelling

The remaining parts of this chapter will consider Bayesian models of language learning. This section introduces the basic concepts behind these models. Bayes' law provides a way to assign probabilities to hypotheses given evidence that is observed and prior knowledge of the situation. Bayesian language learners calculate the rational language (or 'hypothesis') to adopt given the languages they observe being spoken to them and a learning bias. This gives a simple way of modelling a learner who learns from data but also has a parameterisable 'innate' bias towards learning certain kinds of language. As an example, consider a learner trying to establish the sentence order of syntactic items such as subject, verb and object. The possible 'hypotheses' include SVO, SOV, VSO, VOS, OSV and OVS. A learner assigns a probability to each hypothesis, and then selects a hypothesis to use based on these probabilities (there are different ways to do this, see section 5.4.4). The overall probability of a hypothesis (the posterior probability) is calculated from two sources - the learning bias (prior probability) and the observed data (likelihood). These are described below.

5.4.1 Likelihood

The likelihood of a hypothesis is simply the probability that it would generate the observed data. So, for instance, observing many sentences that agreed with SVO order would mean the likelihood of the SVO hypothesis would be high and the likelihood of the VSO hypothesis would be low.

5.4.2 Prior

The prior probability or prior bias is a probability that is assigned to a hypothesis independently of the data. For example, a learner may assign higher prior probabilities to languages that have subjects at the start of a sentence due to a processing advantage given typical semantic processing priorities (see Diamond, 1991, 143). Griffiths and Kalish suggest the following interpretation of the prior:

“The standard interpretation of the prior ... as representing the extent to which the learner believes in a hypothesis before seeing any data is perhaps not the best way to understand the role that it plays under this view of language acquisition. The prior is better seen as determining the amount of evidence that a learner would need to see in order to adopt a particular language. Thinking of the prior as expressing the amount of evidence a learner would need in order to choose a particular language makes it clear how it can encode the biases of learners: only hypotheses with positive prior probability will enter into consideration, and hypotheses with higher prior probabilities are easier to learn (requiring less evidence, and ultimately less data).”

(Griffiths and Kalish, 2007, p.450)

5.4.3 Posterior

A total probability for a given hypothesis, called the posterior probability, is calculated by combining the likelihood and the prior probabilities using Bayes' law (Bayes and Price, 1763). This gives the rational posterior probability for a hypothesis given the likelihood and prior probabilities. The probabilities for all hypotheses can be calculated to give a probability distribution over languages.

5.4.3.1 A non-linguistic example

To clarify the Bayesian model, here's an intuitive example. Imagine that I have a die and you have to guess how many sides it has. The possibilities of the die having 1 side, 2 sides, 3 sides and so on are your 'hypotheses'. I roll the die behind a screen so you can't see. I then tell you the value on the face (the data). Imagine I roll the die three times and the highest number you observe is 6. You're likely to think that I have a 6 sided die. However, imagine instead that I rolled the die three times and the highest number you observed was 3. You might still want to guess I have a 6 sided die, because you know those are the most common type of dice. This knowledge represents your prior bias. Now imagine that I continue to roll the dice a million times and still the highest value is only 3. At this point, you might be more willing to guess that I really do have a 3 sided die. The posterior probability captures the balance between your prior bias and the data you observe.

5.4.4 Selecting a hypothesis

After assigning probabilities to each hypothesis, a Bayesian language learner has to choose a hypothesis with which to produce data for the next generation. Griffiths and Kalish (2007) identify two types of learning algorithms. Maximum a posteriori learners (MAP) choose the hypothesis to adopt for production that has the maximum posterior probability. Sampling learners (SAM) choose a hypothesis by sampling the hypotheses in proportion to their posterior probabilities (i.e. they may adopt any hypothesis, but are most likely to adopt the one with

the highest posterior probability). Learning algorithms can be implemented that interpolate between these two extremes (Kirby et al., 2007).

Dediu (2009) reviews previous work on the differences between these learning algorithms when implemented in an iterated learning framework. This involves learners learning a language, then producing input for the next generation. The stable distribution of languages in chains of SAM learners will converge to the prior (Griffiths and Kalish, 2007). The ‘distribution of languages’ is the probability of a given learner selecting each language in a given generation. So, the ‘stable distribution of languages’ is the value of this probability distribution after a chain has been run for many generations. In the SAM case, the stable distribution is equal to individuals’ prior probability distribution over languages. This is due to an iterative effect of the prior distribution favouring certain languages which are then produced with a probability dependent on that prior distribution.

Griffiths and Kalish (2007) show that the output of SAM learners can be used to estimate their prior probability distribution, but the same is not necessarily true of MAP learners. The stable probability distribution over languages tends to exaggerate the individual prior probability distribution. For example, consider the following (rather extreme) prior probabilities: SOV: 60%, SVO: 40% and all others: zero probability. Initial learners may produce a range of sentence types. After many generations, SAM learners may will be producing SOV 60% of the time and SVO 40% of the time. On the other hand, MAP learners will only be producing SOV sentences.

Dediu (2009) looks at populations of mixed learner types. Populations of SAM learners behave like single chains of SAM agents (converging to the prior) and populations of MAP learners behave like single chains of MAP agents (one hypothesis dominates). However, mixed populations of SAM and MAP learners behave more like single chains of SAM learners. Dediu then added complex social dynamics. This involved many populations in a structured space with overlapping generations, mating, learning primarily from the mother, a small amount of learning from neighbouring populations and migration between populations. Here, the stable distributions converge to the individual’s prior, even populations of only MAP learners.

Smith and Kirby (2008a) show that the MAP algorithm is the evolutionarily stable strategy. They set up a population of SAM and MAP learners where individuals reproduce, passing on their learning strategy to their offspring. The probability of an individual reproducing was tied to ‘communicative accuracy’, that is, the proportion of the population with the same hypothesis as that individual (see section 5.8 for more on measures of communicative accuracy and bilingualism). A population of SAM learners could always be invaded by a population of MAP learners. This was because, given the same data, MAP learners will converge on the same hypothesis while SAM learners may choose different

hypotheses. This means that the ‘communicative accuracy’, and therefore fitness, of the SAM learners was lower. Based on this, Kirby et al. (2012) conclude that “sampling is a bad strategy for coordination problems like language”.

However, Ferdinand and Zuidema (2009) demonstrate that this result is due to a symmetrical hypothesis space. With an asymmetric hypothesis space, SAM learners may be more stable because they have fewer hypotheses to converge upon. If the structure of the hypothesis space means that there are multiple preferred hypotheses, then adopting a single language may leave an agent less likely to be able to communicate with the whole population. Put another way, if agents receive the same data, then MAP learners will converge on the same hypothesis. However, if agents receive different samples of data or data from different sources, then MAP learners might choose opposing hypotheses, while samplers would be more likely to overlap.

The top down model presented in the next chapter allows agents to adopt multiple languages. This leads to an asymmetric hypothesis space which has an effect on the stability of different approaches to learning. However, it remains to be tested whether SAM and MAP algorithms are better adapted for different hypothesis space structures.

5.5 Modelling cultural evolution in a rational framework

A Bayesian framework has been used to model the interaction of individual learning, cultural transmission and individual learning biases (Griffiths and Kalish, 2007; Kirby et al., 2007; Niyogi and Berwick, 2009; Ferdinand and Zuidema, 2009; Smith, 2009; Burkett and Griffiths, 2010; Smith and Thompson, 2012; Thompson, 2012). In these models, generations of learners are exposed to linguistic data and adopt a language type to use themselves. They do this by considering the data they observe and a prior bias over the language types they expect to observe. The prior bias is a preference for a language type that is independent from the data. The learners produce data from this language type for the next generation and the process iterates.

There are two opposing results regarding the link between the distribution of linguistic features and the learning biases of individuals. Griffiths and Kalish (2007) show that there are some assumptions that lead to a transparent link between individual biases and population-level phenomena. That is, the linguistic distribution can be used to make inferences about the learning biases of individuals. However, models using other assumptions show that weak cognitive biases can be amplified by cultural transmission into strong linguistic universals (Kirby et al., 2007; Smith, 2009). In this case, the population-level phenomena are not

isomorphic to the individual biases of learners. This goes against a strong nativist assumption that there is a direct link between the distribution of linguistic typologies in the world and the innate learning biases of individuals (Kirby et al., 2012; Thompson et al., 2012). The recent debate has focussed on whether the link between individual biases and population-level phenomena remains opaque under a range of conditions. This has meant extending the model from single agents learning single languages to include multiple learners in each generation and the ability for an individual to adopt multiple languages.

In order to obtain results for more complex models (e.g. multiple speakers), models make simplifying assumptions about certain aspects (e.g. languages are monolithic and discreet). I will argue that researchers have been biased towards making assumptions that limit the validity of these models relating to bilingualism.

5.5.1 Summary of Bayesian models

This section provides a short summary of the history of cultural evolution with Bayesian language learners. A more detailed description of a more advanced model is given in the next section.

The first Bayesian language learning models included agents who could only adopt single language types (e.g. Griffiths and Kalish, 2007). The initial Bayesian models found opposing results with regards to how individual biases and population-level phenomena are linked. Griffiths and Kalish (2007) showed that the distribution of linguistic types converged to the prior distribution over those types ('convergence'). For instance, in a model with two language types, if the prior bias for type A was 20% higher than type B, then eventually the distribution of language types (either over an infinitely long chain of single learners or in a single generation with a large population) would include 20% more speakers of type A than type B. However, a different result was obtained with MAP learners (see section 5.4.4): the process of cultural transmission could exaggerate the eventual distribution of language types, so that the proportion of the language type that was favoured by the prior bias was higher than the prior probability assigned to it ('non-convergence', Griffiths and Kalish, 2007; Kirby et al., 2007). That is, the cultural transmission process complicated the link between individual biases and population-level phenomena.

Niyogi and Berwick (2009) criticised these models on the basis that only learning from one individual "doesn't embrace the full darwinian picture" (Niyogi and Berwick, 2009, 10124). Smith (2009) modelled learning from multiple teachers, but where learners still adopt only one language. In this case, one of the linguistic types tended to dominate (non-convergence), according to the bias, the amount of data observed, the number of teachers and the initial distribution of linguistic types.

Ferdinand and Zuidema (2009) demonstrate that convergence or non-convergence is sensitive to population size, population structure (see also Stadler, 2009), prior bias homogeneity and the structure of the hypothesis space⁵. They conclude that maintaining rationality in models with complex social structures is difficult. Noting that Smith’s model violates the rationality assumption, Burkett and Griffiths (2010) find a way of letting agents rationally learn multiple languages from multiple teachers. This introduces a prior bias over the amount the linguistic homogeneity of an agent’s input. Extensions of this model allow the strength of this bias to be learned (Smith and Thompson, 2012) or evolved (Thompson et al., 2012). The latter suggests that expecting low variation in the input (‘monolingualism’) is the solution that is most likely to evolve.

The next section presents the model of Burkett and Griffiths (2010) in detail and discusses whether it is a good model of bilingualism. The sections after that discuss the extensions to this model and demonstrate that they make implicit monolingual assumptions. I will demonstrate that considering bilingualism changes the nature of the debate about how individual biases and population-level phenomena are linked. The focus on the debate has shifted from assumptions about separate learning biases to assumptions about expectations about diversity in the input.

5.6 Learning multiple languages from multiple teachers

Burkett and Griffiths (2010) present a model where agents learn “multiple complete linguistic systems” from “truly divergent inputs” (Burkett and Griffiths, 2010, p. 60). That is, learners take into account that the data they receive may be generated by different speakers who may speak more than one language. Below I summarise the model.

⁵They implement two population structures: monadic, with a single chain of learners, as in previous models and polyadic, where the same data is passed to many individuals at each generation, analysed separately, then the data generated by each individual is pooled for the next generation. They find that the stable distribution for samplers in a polyadic chain does not converge to the prior. Furthermore, while samplers converge to the prior with symmetrical hypothesis spaces (as in other models), they diverge from the prior as the structure of the hypothesis space becomes more asymmetrical (the model in the next chapter also has an asymmetrical hypothesis space). They also find differences between samplers and maximisers in populations where individuals have different prior biases. The stable distribution of heterogeneous samplers converges to the average of their prior biases. However, the relationship between the stable distribution for heterogeneous maximisers and their prior biases is more complex (see Ferdinand and Zuidema, 2008b). Ferdinand and Zuidema (2009) suggest that MAP learners may not always be the evolutionarily stable strategy (in contrast with the results of Smith and Kirby, 2008b, see section 5.4.4).

5.6.0.1 Hypothesis space

The data that the agents observe consists of words which are produced by a language with a certain probability. An agent's hypothesis is then a probability distribution over languages. For example, in a situation where there are two possible languages, L_1 and L_2 , an agent could have a hypothesis h where $p(L_1|h) = 0.6$ and $p(L_2|h) = 0.4$. This is a continuous hypothesis space. A 'monolingual' hypothesis would be one where the whole distribution is assigned to one language (e.g. $p(L_1|h) = 1$ and $p(L_2|h) = 0$). By contrast, then, a 'balanced bilingual' would have a hypothesis with a uniform distribution ($p(L_1|h) = 0.5$ and $p(L_2|h) = 0.5$).

5.6.0.2 Likelihood

The likelihood of a hypothesis producing the data that a learner observes is calculated according to the process of production. Agents produce data by, first, selecting a language based on the distribution in their hypothesis. In the example above, the agent would choose language L_1 60% of the time (where $p(L_1) = 0.6$ and $p(L_2) = 0.4$). Then, words are produced given the selected language. For instance, in a system with w_1 and w_2 , the probability of L_1 producing w_1 is ϵ (set close to 1) and the probability of L_1 producing w_2 is $1-\epsilon$ (therefore, very low). The opposite is true for language L_2 .

5.6.0.3 Prior

The prior bias in the model is proportional to two factors: First, a bias over the languages, which we'll call β , like in the other Bayesian models above. Secondly, the concentration parameter α , which represents how often a learner expects to hear new languages. This can range from expecting a single language in the entire population (as α nears zero) to expecting each teacher to speak a different language (as α nears infinity). This method provides a general solution to cultural transmission in populations, thus maintaining rationality (see previous section).

5.6.0.4 Posterior

A learner observes utterances and calculates the most likely distribution over languages given the likelihood of each hypothesis and their prior bias according to Bayes' law.

5.6.0.5 Iteration

The agents in this model use a sampling algorithm to select a hypothesis. They then produce data according to this hypothesis, as described in section 5.6.0.2. This data becomes part of the input for the next generation of learners. This process of exposure to data, induction of a hypothesis and production of data is iterated for many generations. The change in proportions of hypotheses over

generation can be tracked. From these dynamics, the stable distribution of hypotheses can be estimated. This represents the probability of a given generation adopting each hypothesis.

5.6.1 Results

The results show that with high α , the distribution of languages in the population comes to reflect the prior distribution (convergence, in line with the results of Griffiths and Kalish, 2007). In this situation, learners effectively expect each teacher to speak a different language. This means that the dynamics are more like a chain of single learners. That is, path of a single language back in time through a transmission chain of learners does not diverge. However, with a low α a single language type tends to dominate, depending on prior biases and initial data conditions (non-convergence, in line with the results of Smith, 2009). In this situation, agents assume that all the speakers in a population speak the same language, as in a monolingual society. That is, the α parameter defines a continuum in the model dynamics which range from convergence to the prior to sensitivity to initial conditions (non-convergence).

Therefore, the two seemingly opposing results about how individual biases and population-level phenomena are related (convergence to the prior and non-convergence) can in fact be described as extreme ends of a single continuum. This continuum is realised by taking into account multiple languages, multiple teachers and the structure of the population. Arguably, by considering bilingualism, the nature of the debate has changed. The transparency of the relationship between individual biases and population-level phenomena now depends on the expectations that learners have about the linguistic homogeneity of their input.

Some obvious questions arise: Is there an analogue for this expectation in human infants? If so, what is the default expectation, or is it a bias that can flexibly adapt to the situation? Some extensions of the model have produced results relevant to these questions (see sections 5.7 and 5.8). First, however, I consider whether this model is a good model of bilingualism.

5.6.2 Representing bilingualism

Burkett and Griffiths (2010)'s model allows agents to adopt multiple culturally transmitted features from multiple speakers. This section considers whether it is a good model of bilingualism by comparing it to the list of requirements for models of bilingualism developed in chapter 3.

The first requirement concerned the representation of languages. The purpose of Burkett and Griffiths (2010)'s model was not to model bilingualism, but to model how individual biases and population-level distributions were linked. However,

the results relating to this issue rely on an assumption of the ability to divide the linguistic variation into discrete languages:

“Intuitively, we expect that if learners are able to appropriately separate their input into distinct languages, then the learning dynamics will resemble those from the single teacher setting.”

(Burkett and Griffiths, 2010, p. 60)

Whether discrete languages are concrete concepts is a complex issue, discussed in the previous chapter. As Griffiths and Kalish (2007) argue, the processes in top-down models do not need to implement real processes of learning to be valid. Infants certainly can differentiate between languages in their input (see section 3.9 in the last chapter). However, the assumption in the model is that learners can divide linguistic variation into the same whole languages as their cultural parents. That is, there is a fixed set of possible discrete languages that is constant across all time. Put another way, a word belonging to language 1 in the last generation cannot be interpreted as belonging to language 2 in the next generation. Chapter 3 argued against this assumption, suggesting that whole linguistic systems are defined dynamically by use. Although the learners observe low-level features, the linguistic space of languages is static.

There are two possible solutions to this. First, the parameter that dictates how likely a word is to be produced by a given language (ϵ) could also be induced, rather than set to a fixed value. This would allow the mapping between words and languages to change. Secondly, Reali and Griffiths (2010) find a solution for a similar model for a hypothesis space with infinitely many languages. This means that subsequent generations are unlikely to produce exactly the same languages.

However, there are further problems connected with these solutions: the division of languages cannot be conditioned on the identity of the teachers. Agents receive their data from multiple teachers, but do not take into consideration how the variation is distributed across those teachers. That is, the identity of speakers is not observable, and agents induce how homogeneous the population is, not how homogeneous each individual is. A high value of α actually means that learners expect every teacher to speak all possible languages (Smith and Thompson, 2012). This fails the second requirement for a bilingual model that individuals should be represented explicitly. The next chapter suggests a model where agents can identify speakers (although with a simpler model of transmission).

The final requirements are also not met. Agents have the same probability of receiving data from all teachers, so the social structures are not complex⁶. Neither are the social structures dynamic. While the α parameter does have implications for assumptions about social structures, it cannot change in this model (see the

⁶Although Smith (2009) does look at the effects of the number of cultural parents.

sections below for models where it can change). Given the aim of the model, and given that the dynamics of this kind of system are not well understood, this is a reasonable assumption. Also, as Ferdinand and Zuidema (2009) suggest, defining more complex social structures can lead to much more complex models, meaning tractability can become an issue. Researchers can therefore be biased against exploring complex social structures to begin with. In chapter 7 I construct a bottom up model of cultural transmission that allows complex, dynamic social structures from the start. However, rather than seeing the bottom up model as opposing the top down models, it is hoped that the bottom up model makes clear the features that are relevant for a model of bilingual cultural transmission. This could help with the extension of the lineage of top down models.

However, for the purposes of this chapter, top down models of cultural transmission appear to have assumptions that limit the ability to explore bilingualism, even when they model the transmission of multiple languages.

5.7 Learning the prior bias

Burkett and Griffiths's model specified an innate bias for the amount of linguistic variation to expect in the input. What is the most likely setting of this bias for human learners? Smith and Thompson (2012) explore this question through two extensions of Burkett and Griffiths's model: One where agents learn α and one where α evolves. This section and the next describe these models.

Smith and Thompson (2012) use iterated Bayesian language learners who adopt hypotheses which are distributions over languages, as in Burkett and Griffiths's model above. In the first extension, agents estimate the homogeneity of the population (the value of α) at the same time as estimating their hypothesis over languages. This means that an agent's linguistic experience affects their prior biases, in line with experimental work showing that bilinguals make different assumptions about the variation in their input from monolinguals (e.g. Byers-Heinlein and Werker, 2009a; Healey and Skarabela, 2009; Houston-Price et al., 2010; Merriman and Kutlesic, 1993; Pruden et al., 2006; Hirsh-Pasek et al., 2000; Brojde et al., 2012; Healey and Skarabela, 2009; Kovács and Mehler, 2009; Costa and Santesteban, 2004, see section 5.9.3 on page 82).

An agent adopts an α value using Bayesian inference, based on the amount of variation they observe in their input, and a prior bias on this observation. This requires a prior over α that specifies an agent's expectations about their expectations about linguistic homogeneity of a population. This higher-order prior can be uniform, so the agent adopts a value of α based on the data alone, or it can be skewed towards adopting a particular value of α . We'll call the peak of this

skew δ (the value they are most biased towards adopting)⁷. The results show that in populations with low δ , α converges on δ . Also, with low δ the stationary distribution of language types is sensitive to initial conditions (non-convergence, in line with Smith, 2009). That is, there is convergence to the prior over α , but not to the prior over languages (β). The result is the opposite for high values of δ . Higher δ leads to convergence to the prior over β (convergence, in line with Griffiths and Kalish, 2007), but an exaggeration of α .

Experimental evidence shows that infants' expectations about the homogeneity of their input is flexible and adapts to their experience (see section 5.9.3). In this case, infants raised in a monolingual community may expect little variation in their input, like agents with low α , while infants raised in a bilingual community are like agents with high α . The model therefore suggests that there may be qualitative differences between how prior biases are linked to population-level distributions in monolingual and bilingual communities.

5.8 Evolving the prior bias

If the expectation about the linguistic homogeneity of a population is innately specified, then the precise value might evolve over time. Burkett and Griffiths (2010)'s model has been extended so that α is 'biologically' specified and evolves under a selection for communication (Thompson et al., 2012; Smith and Thompson, 2012; Thompson, 2012).

This is implemented in the following way. Each agent has a string of zeros and ones representing alleles in their genome. The prior bias is the proportion of allele types. This is under selection for communication: individuals inherit their genes from a single parent, with a small probability of point mutation. An individual's probability of reproducing is related to how accurately it could communicate with others in the population. This is assumed to be the proportion of individuals in the population who speak the same language.

Smith and Thompson (2012) demonstrate that there is strong selection for low values of α . That is, populations typically evolve to expect low variation in their input. Recall that Burkett and Griffiths (2010) find that low α leads to non-convergence. However, this result is dependent on the measure of communicative success that determines an agent's fitness. Thompson (2010) demonstrates that with different assumptions, high α can evolve (see the section 5.8.1 below for a criticism of these measures).

⁷This is labelled θ in Smith and Thompson, 2012, but conflicts with the use of θ in other models, so I have re-labelled it here for consistency.

5.8.1 Measures of fitness based on communicative accuracy

The fitness of the agents in the evolutionary models (Thompson et al., 2012; Smith and Thompson, 2012; Thompson, 2012; Kirby et al., 2012) is linked to communicative accuracy. Communicative accuracy is the probability that two agents are able to exchange signals that align with the same meaning for each agent. There are two problems with the implementation of this measure in (Kirby et al., 2012). First, they assume that communicative accuracy can be measured directly from their hypotheses. This makes assumptions about communication that favour monolingualism. Secondly, a straightforward measure of communicative accuracy is not necessarily a good index of reproductive fitness. Some alternative measures are suggested that highlight this point.

A measure of communicative accuracy is fairly straightforward for agents who adopt a single, discrete hypothesis (although assumptions could be made about weighted similarity as in Greenberg, 1956). However, agents in the ‘multilingual’ model adopt a distribution over languages. Thompson (2010) defines two possible measurements of communicative accuracy for this case. Here I describe those measures and demonstrate that the measure chosen in Kirby et al. (2012) actually fits a narrow definition of bilingualism.

Type 1 (used in Smith and Thompson, 2012; Kirby et al., 2012), “Monolingual”: Fitness is proportional to the probability that two agents will produce the same language at a given time. Formally, if $P(x_h)$ is the probability that agent x will sample language h , then the communicative accuracy CA between individuals A and B is

$$CA(A, B) = \sum_{h \in H} P(A_h) \cdot P(B_h) \quad (5.1)$$

Therefore, the fittest pair are the ones that will always produce the same, single language. The least fit pair will be the ones that always produce different languages. Pairs that produce languages with the same frequency, but with no strong skew towards one language will receive a median fitness. This measure has been used in other models (Nowak et al., 2001; Briscoe, 2000).

However, the measure privileges ‘monolingualism’. If two agents both only speak L_0 , they receive a communicative accuracy (CA) of 1.0. If both agents speak language L_0 50% of the time, then $CA = 0.5$. If two agents speak L_0 and L_1 40% and 60% of the time respectively, $CA = 0.48$. The conception of language here is as a monolithic, discrete entity, so that two language types are completely mutually unintelligible. The measurement is also equivalent to the following interpretation: Given agent A’s typical proportion of language types, calculate the proportion that agent B will understand, assuming that agent B comprehends L_0 in proportion to θ and L_1 in proportion to $1 - \theta$. This means that the θ value is taken as a measurement of comprehension, too. However, the assumption means

that it is impossible to be fully competent in both languages. This means that an agent with $\theta = 0.5$ is an analogue of a ‘semilingual’ individual (or more accurately, a ‘double semilingual’) who does not have native competence in any language (Bloomfield, 1927; Hansegård, 1968; Cummins, 1976). Martin-Jones and Romaine (1986) characterises this view as imagining discrete languages as containers with an ideal finite capacity that are ‘filled’ with input. The containers are assumed to be universal and fixed (see chapter 3 for arguments against this). There is an implicit assumption that ideal monolingualism is the ‘normal’ course of language development. This view of competence has been criticised as not being supported by linguistic evidence (Skutnabb-Kangas, 1981; Martin-Jones and Romaine, 1986). Part of the problem here is that production and comprehension are assumed to be completely dependent. Before discussing this further, I will present the second measure (described in Thompson, 2010).

Type 2, “Parity”: Communicative accuracy is determined by the similarity between agents’ hypotheses. This can be calculated directly from the hypothesis value. If θ_x is the hypothesis of agent x , then the communicative accuracy is just the difference in θ between two agents:

$$CA(A, B) = 1 - |\theta_A - \theta_B| \quad (5.2)$$

So agents who only speak L_0 receive a communicative accuracy of 1, as do agents who only speak L_1 (as in type 1), or two agents who both speak L_0 40% of the time and L_1 60% of the time ($\theta = 0.4$). If one agent speaks L_0 40% of the time and the other speaks L_0 60% of the time, then $CA = 0.8$. This measure means that all homogeneous populations are equally fit. In this case, high α values (expecting high variation) can evolve, but this is still not underpinned by strong innate constraints for specific language types (Thompson, 2010).

The conception of a language with this measurement is slightly different. Individuals understand each other best when they use similar proportions of each language. The measure can be characterised as follows: If agents produce utterances in proportion to their θ , what’s the maximum proportion of utterances that would be the same language type? Put another way, given the typical utterances of two agents, how efficient is the optimal alignment?

The difference between type 1 and type 2 is similar to the distinction drawn in the bottom up model presented in chapter 7 between two different types of intelligibility (see section 7.3.2 on page 7.3.2). One type - comprehensive intelligibility - measures the similarity of agents’ utterances as an analogue of the proportion of utterances that one learner typically produces that another understands. This is broadly analogous to communicative accuracy type 1 here. Another type - functional intelligibility - measures the proportion of utterances that interlocutors understand when they design their utterances for each other. Communicative accuracy 2 above is similar to functional intelligibility in the sense that agents

‘accommodate’ each other by aligning their utterances. In the bottom up model, the difference between these two measures is taken as an index of bilingualism.

However, successful communication is usually thought of as involving the production and *comprehension* of an utterance. There are no assumptions about comprehension in the model. Agents ‘know’ that both types languages are possible, and may be exposed to both languages, but they adopt a distribution over the *production* of these languages. For real learners, the proportions of production and the level of comprehension of a language can be independent. For instance, children may develop a passive competence in a minority language (Lincoln, 1979; Baker and Jones, 1998, p. 495; Baker, 2011, p. 101-102), and ‘non-reciprocal’ language use is common in trade situations (see Croft (2003)). The literature on minority language acquisition also demonstrates that the proportion of languages in a child’s input is not simply related to the resulting comprehension in each language (e.g. De Houwer, 2007). If it were, then bilinguals with native competence in both languages (see Peal and Lambert, 1962; Butler and Hakuta, 2004; Grosjean, 1982; Myers Scotton, 2006) could not exist. Since there’s no division in the model between production and comprehension, conceptions of bilingualism can be difficult to realise.

5.8.2 Alternative measures of fitness

The two measures above are not the only possibilities, however. Below I list some other possible measures of fitness that could be explored⁸.

Type 3, “Linguistic Exogamy”: Assume that agents preferentially mate with other agents who are most linguistically different. This models linguistic exogamy: communities where marriage is restricted to members of different linguistic communities (e.g. Jacobs, 1937; Sorensen, 1967; Monod, 1970; Jackson, 1983). This practice often maintains multilingualism (see also Hill, 1978, p.13-16). The appropriate fitness payoff function FP is therefore the inverse of the communicative accuracy function:

$$FP(A, B) = 1 - \sum_{h \in H} P(A_h) \cdot P(B_h) \quad (5.3)$$

Type 4, “Bilingual”: Assume that agents who are bilingual receive the highest fitness. Prestigious bilingualism is attested in many communities, and is often linked with the power to communicate between groups (see e.g. Laycock, 1985; De Mejía, 2002; Hning et al., 2012; Nettle, 1999b). We can reflect this in the measurement by weighting the fitness payoff by how close the agent’s θ is to the

⁸To bolster the arguments in chapter 8, these suggested measures and implications emerged from actually coding and visualising the functions.

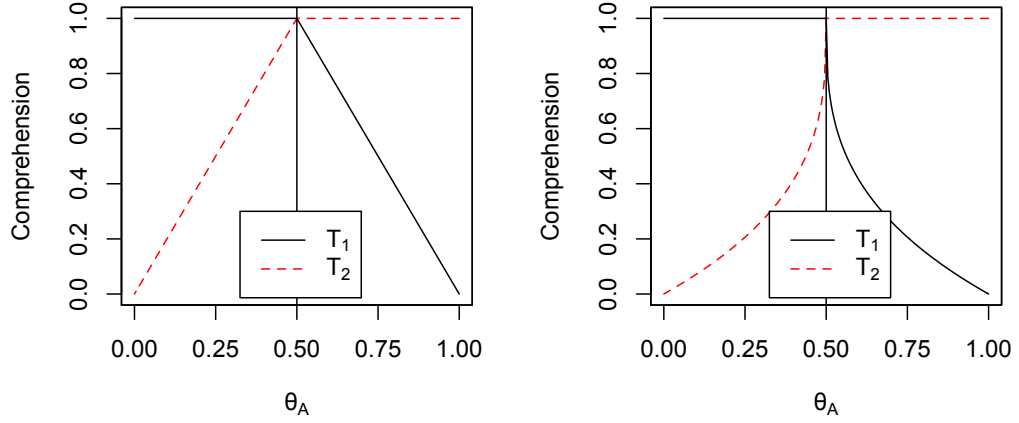


Figure 5.1: The relationship between an agent’s hypothesis θ and its comprehension of language types L_0 and L_1 for the ‘dominant language’ measure of communicative success, with a linear (left, $\gamma = 1$) and exponential (right, $\gamma = \frac{1}{3}$) function.

midpoint:

$$FP(A, B) = 2 \sum_{i \in H} P(A_i) \cdot P(B_i) \cdot |\theta_A - 0.5| \cdot |\theta_B - 0.5| \quad (5.4)$$

Type 5 “Dominant Language”: Assume that a speaker always understands their dominant language, and understands their non-dominant language in proportion to the balance of their θ , according to some function. The level of comprehension of language type L_i (where $L_i \in 0, 1$) by agent x is:

$$Comp(x, L_i) = \begin{cases} 1 & \text{if } |\theta - L_i| < 0.5 \\ G(|\theta - L_i|) & \text{otherwise} \end{cases} \quad (5.5)$$

Where

$$G(x) \propto \left(\frac{1}{x}\right)^\gamma \quad (5.6)$$

For instance, $G(x)$ can return values so that the comprehension of the weaker language is linearly ($\gamma = 1$) or exponentially (e.g. $\gamma = \frac{1}{2}$) related to the balance of θ (see figure 5.1)⁹. The variable γ , therefore, describes how difficult a second language is to learn, fitting one of the premises of the bilingual paradox. An exponential function reflects the idea that native competence in a second language is hard to obtain. The fitness payoff is therefore calculated as

⁹The ‘parity’ measure is a special case of the ‘dominant language’ measure where $G(x)$ is constant.

$$FP(A, B) = \sum_{i \in H} \sum_{j \in H} P(A_i) P(B_j) \text{Comp}(A, L_j) \text{Comp}(B, L_i) \quad (5.7)$$

Figure 5.2 shows how the communicative accuracy between two agents varies as a function of their hypotheses for the different measures described above. The ‘monolingual’ and ‘parity’ measures lead to different evolutionary stable values of the expectation of the amount of variance in the input (Thompson, 2010), so the other measures described here should also have an impact on the range of α (although it is unlikely that this would lead to strong innate biases either).

Furthermore, in the real world, it is probable that the way communicative success and fitness are related will change over time. For instance, the prestige of bilingualism would depend on the linguistic variation being conditioned on social variables, otherwise there would be no social advantage in being bilingual. This situation is unlikely to come about unless the society is stratified. A model that considered the evolution of a bias for linguistic variation (e.g. Smith and Thompson, 2012) should therefore also consider dynamic social structures.

In this section, I have shown that the choice of fitness function in the model of cultural evolution above had an implicit monolingual bias. A careful consideration of bilingualism revealed more parameters that could be explored in the model.

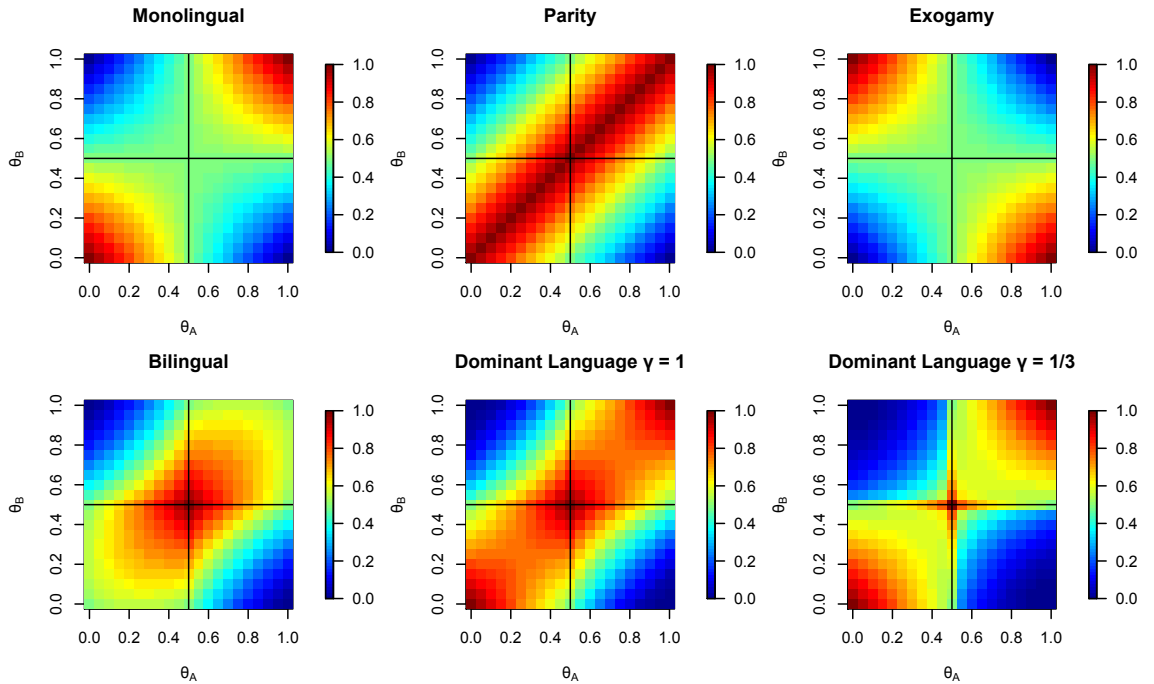


Figure 5.2: The fitness payoff between two agents with different hypotheses (θ values) for different fitness payoff functions. Warmer colours represent higher fitness payoff.

5.8.2.1 Preliminary results

In order to get some idea of the impact of different fitness measures, Smith and Thompson (2012)'s model which includes an evolving α parameter was run with the dominant language fitness metric¹⁰. The 'bilingual paradox' questioned why bilingualism is prevalent if acquiring two languages is difficult. This model allows us to manipulate the difficulty of learning a second language (the γ parameter) and observe the linguistic diversity and expectations about linguistic diversity that emerge.

The model was set up so that individuals had a strong bias towards low values of α (gamma distribution shape parameters = (1,1)) and a moderate bias towards language L_1 ($\beta = 0.6$). A population of 100 individuals evolved for 100 generations and the final values of α and the final mean value of hypotheses in the population were recorded. Individuals reproduced according to the 'dominant language' fitness metric. Simulations were run for a range of the γ parameter of the fitness metric, which dictates how easy a second language is to acquire.

Figure 5.3 shows the results. As γ increases, there is a qualitative shift in the results of the simulations. With $\gamma < 0.5$ (learning a second language is hard), low α evolves ('monolingual' expectation, convergence on the prior) and the distribution of languages is exaggerated (L_1 comes to dominate, non-convergence). However, with $\gamma > 0.5$ (learning a second language is easier), high α evolves ('bilingual' expectation, non-convergence) and the distribution of languages converges to the prior.

There are two interesting observations to be made. First, unlike Smith and Thompson (2012)'s results, high values of α can evolve in some situations. Secondly, Smith and Thompson (2012) found that varying the prior over α interpolated between the convergence and non-convergence results in the model where α was learned. The results above suggest that the same interpolation can be achieved in the evolving α model by keeping the priors fixed and manipulating the part of the fitness metric that controls how easy a second language is to learn. That is, the communicative fitness metric is an important part of the debate about the relationship between individual biases and population-level phenomena. The results suggest that 'bilingualism' can emerge even if learning a second language is moderately difficult ($0.5 < \gamma < 1$).

¹⁰The code for the model was supplied by Bill Thompson. The results presented here are preliminary because they are the focus of current investigation by Thompson, Smith and Kirby.

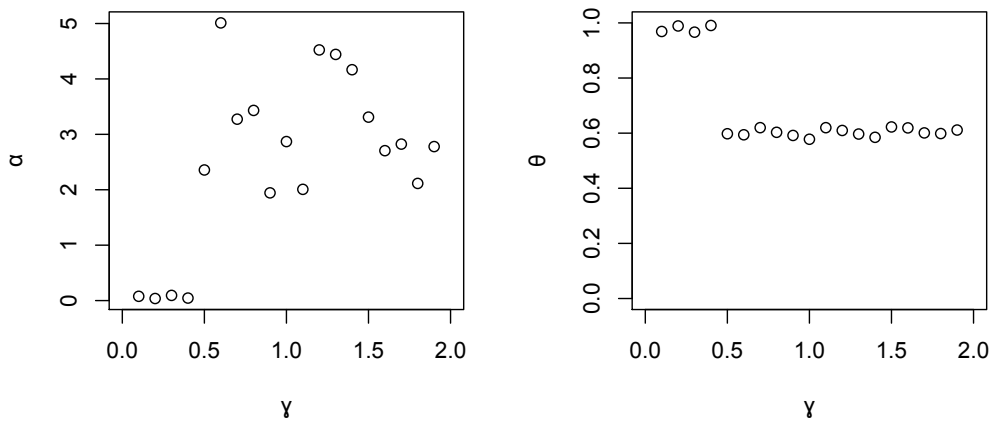


Figure 5.3: Preliminary results for the evolving α model. Left: the final mean values of α in the population as a function of the value of γ (how easy a second language is to learn). The prior over α favoured low values of α . Right: the final mean values of the hypotheses of individuals. The prior over hypotheses was 0.6. Points shown are the mean of the last five generations of a single simulation. As γ increases, there is a qualitative shift in the dynamics of the model around $\gamma = 0.5$.

5.9 The implications of bilingualism for top down models

The sections above reviewed models of cultural evolution that included some element of bilingualism. However, many initial models make simplifying assumptions that undermine bilingualism. The Bayesian models of cultural transmission above have developed to include a fairly complex and domain-specific mechanism that allows a form of bilingualism. However, as Thompson (2012) points out, the inclusion of bilingualism was not motivated by questions about bilingualism, nor because learning multiple languages is more realistic, but to validate the assumptions about rationality when learning from multiple teachers. Researchers have been biased towards making certain assumptions and asking certain questions. These have lead researchers away from modelling bilingualism in a concrete way. Although this paradigm is relatively new, bilingualism has not been a priority in the history of these models.

The next chapter rolls back the developments made above to reveal the differences in research direction that might have been taken if bilingualism was a priority. The first model in the next chapter is essentially a re-implementation of the model from Griffiths and Kalish (2007) and the second model is a re-implementation of the model from Smith (2009). They'll suggest a resolution to the bilingual paradox.

However, this chapter and the last provided some arguments suggesting that the way languages are represented in these models are not valid for a study of bilingualism. In this section I discuss this issue. Section 5.9.1 considers whether the validity of a bilingual interpretation of the models is problematic given the

intended purpose of the models. Section 5.9.2 discusses whether bilingualism can make a qualitative difference to the to what evolves in this kind of model. Section 5.9.3 argues that flexible learning mechanisms are an important part of language learning, based on studies of bilingualism. They should therefore be part of computational models. These arguments result in a challenge to the top down model approach, discussed in section 5.9.3.2. The main challenge is not necessarily to incorporate the complexities of bilingualism, but to continue to produce solid results with more complex models.

5.9.1 The purpose of top-down models

The top down models above describe a problem where individuals must induce the likely state of another individual according to the behaviour they observe and a bias towards inducing certain states. The models demonstrate the problems with linking population-level phenomena with individual biases, intended to oppose strong nativist assumptions about the origins of the structure of language. However, the models are very abstract and are not realistic when it comes to the representation of linguistic features nor the mechanisms of learning (see McClelland et al., 2010). I have also argued that their representations of languages are not compatible with the theory of bilingualism developed in this thesis. However, does this mean that the models cannot provide insights into bilingualism? The answer depends on whether one sees the models as explanations or as tools. The purpose of these models is not necessarily to explain how language learning works, but to be a tool for exploring the dynamics of cultural transmission. As Griffiths and Kalish argue:

“The Bayesian framework is not supposed to be interpreted as a statement of the mechanistic process by which language acquisition takes place, with learners maintaining a hypothesis space in their heads and updating a distribution over those hypotheses. Rather, it is a computational level analysis (Marr, 1982), as is generally emphasized in rational models of cognition ... , focusing on the abstract computational problem and a method for solving that problem. So long as the actual process underlying language acquisition approximates this solution, our results will have implications for understanding human behavior.”

(Griffiths and Kalish, 2007, p.450-451)

The Bayesian framework gives a formal and precise way of specifying a situation where individuals learn from observation and an ‘innate’ bias, and where behaviour is transmitted culturally. The precise specification of this problem allows the causes of phenomena at the population level to be provably linked to properties of individuals. Rather than models of reality, then, these models can be seen as null models (Reali and Griffiths, 2010). Null models demonstrate the properties we would expect to see without more complex explanations.

For example, Reali and Griffiths (2010) demonstrate how distributions of linguistic variants can change over time by drift (i.e. when no variant is ‘fitter’ than any other). The dynamics of their model exhibit similar properties to real languages such as s-shaped language change curves and power-law relations between a word’s frequency and rank and between a word’s frequency and rate of replacement. They suggest that “by identifying which properties of human languages can be produced by iterated learning alone, we can begin to understand when explanations that appeal to other factors are necessary” (Reali and Griffiths, 2010, p.431).

Without a null model, it is difficult to tell whether a certain mechanism is necessary to explain the phenomena of interest. More complex aspects of learning or linguistic features can then be added to these null models and the effects of these more realistic features can be determined by contrast. Furthermore, null models may be crucial for integrating real data into abstract models (Blythe, 2012).

Another way of looking at these models is as ‘idea models’ (e.g Kirby and Hurford, 1997). That is, the cultural transmission of language underpinned by biological cognitive biases is a complex system. It is difficult to intuit about how the system might work, so computational models like the ones described here are used in order to help think about these kinds of systems. They are not necessarily intended as explanatory models that reflect the way a process actually works. The results of these models can work together with empirical evidence and theoretical work towards an argument (see Scott-Phillips and Kirby, 2010 and Cornish et al., 2009, p. 199).

A crucial question is whether the abstractions made by the model are valid. All models must make abstractions, but these will have an effect on the conclusions. I have argued in chapter 3 that the concept of monolithic, discrete, static languages is not realistic. That is, at an implementational level, language learners are not able to perceive what ‘language’ is being spoken. Instead, the perception of different languages is built up from low-level linguistic variation and social aspects. There is therefore a question about whether it is valid to study computational agents that have biases towards particular ‘languages’. However, the Bayesian models studied here are computational level models, not implementation models. It is less clear if a complaint at the implementation level can affect a computational level model. As I noted in chapter 3, it is certainly possible to identify separate languages in specific contexts, so why not assume that learners eventually become sensitive to the distinctions between the languages they perceive and subsequently use this information?

For certain research questions, such as the extent to which cultural transmission can explain language structure, the abstraction to monolithic, discrete, static linguistic units might be valid. However, different research questions have different

relevant abstractions, so care must be taken when borrowing models that were built for one purpose into a model built for another purpose. Abstractions that are suitable for arguing against a strong nativist position may not be suitable for studying bilingualism in a culturally transmitted framework. When looking at bilingualism, I argue that linguistic divisions that can be described as separate languages must emerge from the model, not be pre-encoded. That is, part of the object of interest for a cultural evolution approach to bilingualism is how divisions between ‘languages’ or ‘mediums’ come about in the first place. However, it is only by comparison with the current Bayesian models of iterated learning that I am able to show that certain abstractions have important implications *for my research question*.

In addition to this, however, I make a slightly different claim. In later chapters I will argue that dynamic social structures are an important part of what drives levels of bilingualism. The current Bayesian models have fixed social structures and agents that are not sensitive to the identity of speakers. I compare these with a bottom-up model that I present in chapter 3 which involves dynamic social structures and agents that use social information. I will show that the conclusions about how population-level phenomena and individual biases are linked can differ depending on the assumptions about these issues. In this case, I am revealing an abstraction that has consequences both for research questions concerning bilingualism and for questions of cultural explanations of language evolution. While the results do not necessarily speak against proponents of cultural transmission, they deserve to be explored.

5.9.2 Populations of languages and populations of learners

Bilingualism involves a certain mapping from a population of linguistic variants to a population of individuals. As chapter 3 showed, even within the field of bilingual language acquisition and sociolinguistics, the understanding of how to specify this mapping is contentious and often based on somewhat arbitrary conventions. In the initial models presented in this chapter, learners can only learn one language. Therefore, there’s no difference between counting the systems of individuals and counting the variants in the whole population. This means that there’s only one way to map the two levels, in which case bilingualism, in the sense understood by this thesis, cannot emerge.

Later models included the ability to adopt multiple languages from multiple teachers, introducing a distinction between the population of variants and the linguistic systems of individuals. Therefore, what has come to light by discussing the models above in terms of bilingualism is a distinction between two levels of language: low-level variants and an individual’s system of variants at a higher level. However, agents in the models above don’t use information on which teacher

produced each utterance, meaning they can't take into account the higher-level systems of individuals.

Evolutionary theory suggests that the dynamics of selection at different levels might be different (Godfrey-Smith, 2009). In a model where learners are biased towards expecting little variation (low α), the distribution of variants can converge to a single variant (everyone speaks one language). In this case there is no difference between the distributions of individuals' systems and the distribution of variants. This means that the fidelity of transmission between generations can be high because an individual's system is simple and the likelihood of moving away from the 'monolingual' hypothesis is small. When learners expect a lot of variation (high α), the distribution of variants reflects the prior bias of learners. However, the distribution over systems may be different (as demonstrated in the model in the next chapter), and the fidelity of transmission of systems between generations may be lower than the transmission of variants. As Godfrey-Smith (2009) argues, a shift in the fidelity of transmission can alter the level of selection in an evolving system.

In section 5.9.3 below I will discuss some evidence that children raised bilingually have different expectations about diversity in their linguistic input. Communities of bilinguals should therefore exhibit different dynamics than communities of monolinguals. In line with this, the models above find different relationships between individual biases and population-level phenomena depending on the expectations about diversity (Burkett and Griffiths, 2010; Smith and Thompson, 2012).

I argue that in order to conduct an investigation into bilingual cultural evolution, properties of both the linguistic population and speaker population must be respected. First, there must only be weak selection between linguistic variants - learning one variant does not exclude the learning of another. Secondly, individuals must be fully represented in the model so that there can be variation in the linguistic systems of teachers that can make a difference to the learning process of learners. Accordingly, the next chapter presents a top-down model where individuals do take into account the identity of the speaker. The chapter after that presents a bottom-up model which allows learners' linguistic systems to be sensitive to the identity of individuals.

5.9.3 Flexible biases

Studies of language acquisition are revealing important aspects of the language learning process that models of cultural evolution should try to capture. For example, in the recent conference on the evolution of language (EvoLang9), Saffran (2012) demonstrates that flexible learning biases are an important aspect language learning. Saffran suggested that we should see learning as a chain of processing where linguistic input is processed by a statistical learning mechanism

that causes some kind of linguistic knowledge as output. For instance, exposure to speech sounds cues infants to the phonetic divisions of the language. This initial knowledge then affects the learning of the transition probabilities between phonemes, which cue infants to word segmentation. This segmentation then affects how labels for words are learned. It was striking that Saffran’s depiction of this process was very close to how evolutionary linguists think about the iterated learning process. However, an element central to Saffran’s approach that has not yet been addressed to a great extent in the field of language evolution is the flexibility of biases. That is, allowing the learning mechanisms to adapt to what is being learned (cf. Smith and Thompson, 2012, see section 5.7). Developmental linguistics studies have been demonstrating the flexibility of learning processes (Merriman and Kutlesic, 1993; Byers-Heinlein and Werker, 2009a; Houston-Price et al., 2010; Pruden et al., 2006; Hirsh-Pasek et al., 2000; Brojde et al., 2012; Healey and Skarabela, 2009; Kovács and Mehler, 2009; Smith et al., 2011). A challenge for top down approaches is to model this complex system.

Researchers find it intuitive that the mechanism that handles learning two languages is the same one that handles three (Grosjean, 1989)¹¹, yet there is still a concern about assuming the same learning mechanisms are used by monolinguals and bilinguals (Grosjean, 1989). The cultural evolutionary approach taken in this thesis emphasises that the differences in the way monolinguals and bilinguals approach learning is mediated by their linguistic experience. That is, while all infants start with the same basic learning mechanisms, they are flexible and adapt to the linguistic and social situations they are placed within. Recent experimental work has demonstrated the extent of this flexibility. For example, infants modulate their use of the mutual exclusivity principle (Merriman and Bowman, 1989; Markman and Wachtel, 1988) in word learning based on their language experience (Merriman and Kutlesic, 1993; Byers-Heinlein and Werker, 2009a; Houston-Price et al., 2010). These findings are being investigated using models, approached from the top-down (Frank et al., 2009) and bottom up (Fazly et al., 2010) perspectives.

Infants pay attention to different cues in word learning at different points in development (Pruden et al., 2006; Hirsh-Pasek et al., 2000) and depending on their experience with languages (Brojde et al., 2012; Healey and Skarabela, 2009). Kovács and Mehler (2009) demonstrate that children exposed to two languages adapt their approach to learning. When exposed to two linguistic patterns, bilingual infants learned both while monolingual infants learned one. This helps bilinguals acquire two languages in the same timeframe as monolinguals acquire one (see Pearson et al., 1993; Werker and Byers-Heinlein, 2008).

¹¹“A multilingual language system is potentially noisier than a bilingual language system, but the mental processes and mechanisms that handle this increased level of noise are presumably no different from those involved in dealing with the extra noise in a bilingual system as compared with a monolingual system.” (De Groot, 2010, p. 2). Although see Hoffmann (2001b) for a discussion of the similarities and differences between bilingual and trilingual acquisition.

Exposure to multiple languages can also cause cognitive advantages in non-linguistic domains (see De Groot, 2010, chapter 7, for a summary). For example, thinking in a second language leads to more rational choices (Keysar et al., 2012) and bilinguals may perceive causal agency differently to monolinguals (Fausey and Boroditsky, 2008, 2011), something that Lucas and Griffiths (2010)’s model of causal relationships might be able to address.

Theories of cultural evolution hypothesise that languages change in a response to learnability and expressivity pressures (Kirby et al., 2008; Cornish et al., 2009). But the studies above demonstrate that this learnability is dynamic and depends on the context of learning. For example, the learnability of a language is affected by the previous experience of the speaker, as is evident from studies that show that a first language influences how a second language is learned (see De Groot, 2010). Changes to a child’s social environment (e.g. going to school) can change the kinds of competition that exist between languages (Hoffmann and Stavans, 2007). Learnability also changes over cultural time: cultural evolution can lead to coadaptation of different features of language, making them easier to learn in combination (McCrohon, 2012).

5.9.3.1 Task Switching

Another domain of flexibility is the response to task switching. Bilinguals with an imbalance in their proficiency (e.g. L2 learners) exhibit an asymmetrical language switch cost (Meuter and Allport, 1999). Switching from the more proficient language into the less proficient language is easier than the other way around (because the more proficient language requires more ‘inhibiting’). This predicts that the balance in proficiency is related to the symmetry in the switching costs (see Calabria et al., 2011; Meuter and Allport, 1999; Costa and Santesteban, 2004). In line with this prediction, switching costs between an L1 and an L2 are symmetrical for highly proficient bilinguals (Costa and Santesteban, 2004; Costa et al., 2006). However, in opposition to the hypothesis, trilinguals who are highly proficient in two languages still don’t exhibit an asymmetrical switch cost between a proficient L1 and a weaker L3 (Costa and Santesteban, 2004; Costa et al., 2006). Switching costs for these highly proficient bilinguals was independent of age of acquisition, proficiency and similarity of the target language. Costa and Santesteban (2004) hypothesise that there is a qualitative difference in the mechanism used by proficient bilinguals. The difference in the mechanisms derives from the way communicative intents become uttered lexical items. Below I summarise two hypotheses about this process

The hypotheses derive from the ‘lexical selection’ framework (Caramazza, 1997; Dell, 1986; Levelt, 1993, 2001; Levelt et al., 1999). Under this framework, production proceeds in two steps: First, a communicative intention activates lexical items, including the target lexical item, but also lexical items that are semantically connected with the target. Secondly, a selection mechanism is needed to

select an item for production, based on their activation. Lexical items from both languages of a bilingual are activated (e.g. Colomé, 2001; Costa et al., 2000, 2003; De Bot, 1992; Gollan and Kroll, 2001). However, it is less clear whether lexical items in the other language compete for selection. Some assume that lexical selection is language specific, so that activated lexical items in the non-target language are not candidates for selection (Costa, 2005; Costa and Caramazza, 1999; Costa et al., 1999; Roelofs, 1998). This is similar to theories suggesting that selection mechanisms are sensitive to other properties of lexical items such as word class (i.e. nouns, verbs etc., Dell, 1986, see Costa and Santesteban, 2004). Other models assume that the selection mechanism is insensitive to the target language, and instead the activation of lexical items is different for target language and non-target language items before selection (La Heij, 2005; Poulisse and Bongaerts, 1994; Green, 1986, 1998; Hermans et al., 1998).

As mentioned above, L2 learners exhibit asymmetrical switching costs, consistent with a language-independent selection mechanism. However, the result for highly proficient bilinguals suggests that they have a language-specific selection mechanism (Costa and Santesteban, 2004, p. 505-506). This would account for the symmetrical switch costs even with weaker languages.

The relevance of these studies on switch costs for this thesis are now discussed. First of all, the apparent qualitative difference between different kinds of bilinguals demonstrates the importance of considering experimental participants with a wide range of linguistic backgrounds. More importantly, while this thesis has debated whether monolithic, discreet languages that are stable across time can have a psychological reality, the results here suggest the picture is even more complicated. All bilinguals exhibit language switching costs, so switching between “language schemas” (Green, 1998) does require some cognitive effort, suggesting that there is some cognitive reality to the division of lexical items according to language. However, the results from the highly proficient bilinguals above suggests that the level of processing at which ‘languages’ are relevant can depend on the individual’s linguistic experience. The qualitative difference above could point to a highly flexible learning mechanism¹². Different populations, then, might have radically different learning and production biases.

This research also has an impact on the debate in this thesis about whether bilingualism requires specific learning mechanisms: Calabria et al. (2011) find that bilinguals with symmetric language switching costs have asymmetric switching costs in non-linguistic domains. They argue that, while general executive control

¹²However, it is also possible that the coarse categories of L2 learner versus highly proficient bilingual could be misleading. Costa and Santesteban (2004) use self-selection to recruit and ask participants to report their age of onset, number of years of regular use and self-assessed proficiency. Perhaps the use of more fine-grained, objective measures, as suggested by the individual differences approach, would reveal that individuals had a quantitative difference in a single mechanism ‘type’ that varied on a continuum from greater use of inhibition to greater use of selection.

processes are involved (Abutalebi et al., 2008), there might be sub-systems of executive control processes that specifically deal with switching between languages (Calabria et al., 2011).

The top down model discussed in this thesis assumed there was a specific mechanism for handling multiple languages while the bottom-up model suggested that a single general learning mechanism underlies bilingual and monolingual language learning. While the importance of social variables could change in the bottom up model, it assumed that the mechanism for learning (multiple) languages was not different from the statistical learning mechanisms in any other domain. In both cases, these assumptions were made partly for convenience. However, the results from Calabria et al. (2011) suggest that at least some individuals exhibit specific mechanisms for dealing with multiple languages. If this is the case, then the top down and bottom up models may represent different ends of a continuum of responses to learning.

5.9.3.2 A challenge to top-down modelling

The studies above suggest that learning biases are flexible and adapt to properties of the input. A challenge to top-down modelling that investigates how individual biases and population-level phenomena are linked is to capture this kind of feature. This might be difficult because of tractability issues. Analytical results are difficult to acquire for the most advanced current models, so the challenge is not necessarily whether the rational approach is valid, nor whether top-down models can provide a good argument against nativism, but whether they can continue to produce solid results for increasingly complex models.

One aspect of this is addressing criticisms that the hypothesis space is assumed to be known in advance of learning. The discussion of the flexibility above suggests that learning processes might be able to radically adapt to relevant cues in the input. So, as well as evaluating hypotheses, learners may need to generate them in the first place - two aspects that may be underpinned by different processes (Sulik, 2012).

Another aspect is modelling sequential learning. Experimental evidence also shows that learning strategies are affected by the order of input. For example, Smith et al. (2011) find that participants' cross-situational word learning strategies are affected by the amount of referential uncertainty, but also whether words are presented in blocks or interleaved with other words. The order of the acquisition of languages may also have an effect on how they are learned (e.g. Dewaele, 1998). Bottom-up sequential learning approaches fit intuitively with this effect while many top-down models traditionally batch-process their input (McClelland et al., 2010; Levy et al., 2009).

However, there are several developments in top down methods that are addressing

these problems. Frank et al. (2009) demonstrate that Bayesian models can successfully learn mappings between words and objects from cross-situational data from real child-directed speech. Incremental Bayesian models have been developed (e.g. Blei et al., 2004; Gomes et al., 2008; Blei et al., 2010), and have been applied to language learning to reproduce real psychological memory effects (Levy et al., 2009; Driesen et al., 2011). It would be interesting to see if it would also model affects of language attrition. Statistical Bayesian methods have been applied to unsupervised learning of linguistic structure (e.g. Goldwater and Griffiths, 2007; Gao and Johnson, 2008), including learning from multiple languages simultaneously (Snyder et al., 2008; Naseem et al., 2009, 2012). In the latter case, simultaneous multilingual input actually improved part-of-speech tagging by up to 53% because multiple cues are available. This appears to argue against the premises of the bilingual paradox.

Methods such as variational Bayesian inference (e.g. Attias, 1999, MacKay, 2003, p. 422-436) allow unknown parameters, hidden variables and model selection, which could capture aspects of cognitive flexibility discussed above (see Driesen et al., 2011). These kinds of models have been applied to natural language tasks such as grammar induction (Kurihara and Sato, 2006). Variational Bayesian methods may also yield results more efficiently than Gibbs sampling, used in the models above (Beal, 2003).

Given these advances in empirical top-down methods that use complex data, there is a great potential for developing cultural evolutionary models that can replicate the kind of flexibility in learning that the bilingualism literature exhibits.

5.10 Conclusion

This chapter has identified a number of models of the cultural evolution of languages that make simplifying assumptions that favour monolingualism. Many models represent languages as monolithic, static and discrete, or make simplifying assumptions about learning mechanisms that make acquiring multiple languages difficult.

The top down models above have developed to represent language learners who can speak multiple languages and have an expectation about the amount of variation in their input. These expectations are a key part of the question of how individual biases and population level phenomena are linked. This is an important question for evolutionary linguistics because it can inform arguments against traditional nativist claims about language. Depending on the assumptions about expectations about variation in the input, some results demonstrate that observing strong linguistic universals in the world does not imply that they are underpinned by strong cognitive biases.

However, the models were not designed to study bilingualism, so caution is required when interpreting their results in terms of bilingualism. I have argued that the expectations about the number of ‘languages’ in the models above might not necessarily map onto real cognitive biases, partly because, as argued in chapter 3, ‘languages’ may not have a psychological reality. The next chapter reconstructs some of the models discussed in this chapter with a focus on bilingualism and demonstrates different results from the ones above. This will reinforce the point that a researcher’s assumptions about bilingualism can affect the conclusions of the models they build and the direction of research.

CHAPTER 6

BAYESIAN BILINGUALISM

“Le langage est source de malentendus.”
Language is the source of misunderstandings

de Saint-Exupéry (1943), Chapter XXI

6.1 Introduction

Previous chapters discussed the idea of the bilingual paradox: we expect one language to be easier to learn than two, yet bilingualism is prevalent in the world. Iterated learning with Bayesian agents, then, seems like a good framework for addressing this question since it is designed to study how individual learning biases (like two languages being more difficult to learn than one) are related to population-level distributions (like the prevalence of bilingualism). This chapter explores Bayesian models of iterated learning that allow bilingualism.

The first model simply lets learners adopt more than one hypothesis, but only contains a single learner at each generation. This model explores the differences between samplers and MAP learners and the problems with measuring bilingualism in this kind of model. The second model allows learners to adopt multiple hypotheses from multiple teachers. Also, learners can observe the identity of their teachers and form hypotheses for each teacher independently. The second model demonstrates how the two aspects of the bilingual paradox can be true at the same time. This is due to an unintuitive, complex link between individual learning and population-level phenomena. Indeed, in the second model, bilingualism is more likely than monolingualism, even when individuals have a bias towards learning just one language.

Although neither of these models meet the criteria for rationality that is addressed by Burkett and Griffiths (2010) (discussed in the last chapter), they do highlight the assumptions that go in to the more complex models. I suggest that, if researchers had prioritised the study of bilingualism after the development of the initial Bayesian iterated models (e.g. Kirby et al., 2007; Griffiths and Kalish, 2007), then the current debates might be different. Particularly, the first model demonstrates that there are many ways to measure the stationary distribution, so whether the bias and the stationary distribution are related in a straightforward way does not present a straightforward dichotomy. The second model also

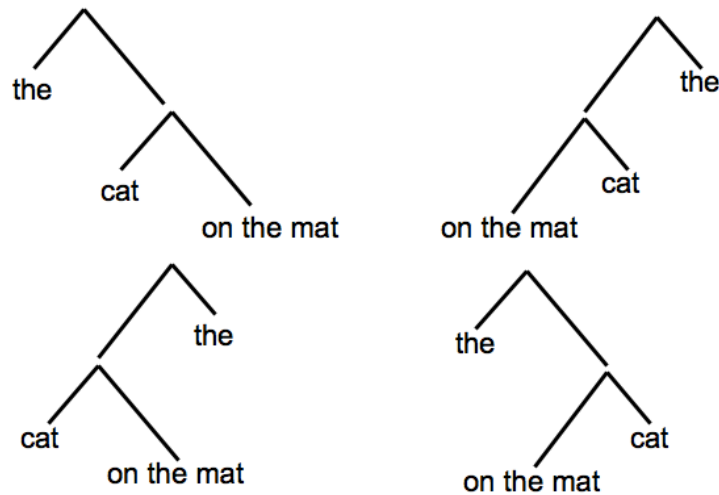


Figure 6.1: Four possible hypotheses in a system with two syntactic variables (NP and prepositions) and two variants (left-branching or right-branching).

highlights the role of population dynamics, speaker identity and the limits of rationality.

6.2 Model 1: Allowing multiple hypotheses

Language learners in initial Bayesian models of cultural evolution were ‘monolingual’: speakers only speak one language and learners assume one language is being spoken (see previous chapter, Kirby et al., 2007; Griffiths and Kalish, 2007). This section looks at what happens to the dynamics of language evolution when a learner can adopt multiple languages. Although Burkett and Griffiths (2010)’s model already addresses this issue, it is a useful exercise. Not only will it demonstrate that the dynamics of language evolution are different with monolingual assumptions, but it will show the extent to which a monolingual approach to the analysis differs from the analysis demanded by the bilingual approach. The differences between samplers and maximisers will become apparent.

6.2.1 Model definition

This section defines model 1. The linguistic system has two variables with two variants each. Imagine, for instance, a system with only determiners, nouns and propositions. Sentences like “The cat on the mat” and “The on the mat cat” are possible. The variants are the two syntactic levels and the variants are right-branching or left branching. Figure 6.1 shows the 4 possible different ‘languages’.

A teacher transmits variants of variables from these languages to a learner. This can be interpreted as the learning hearing “the cat” or “cat on the mat”, for example. The number of variants transmitted represents the ‘bottleneck’ on

learning (here the bottleneck is set to 5 utterances). There is also some noise in transmission - that is, a speaker may ‘accidentally’ produce a variant from another language, or the learner may misperceive a variant as belonging to another language. The learner tries to figure out which hypothesis the teacher is using by Bayesian inference. A hypothesis consists of one or more languages. It then adopts this hypothesis, and uses it to produce variants for the next learner in the chain. There is only one individual at each generation.

Learners have a prior bias towards learning hypotheses that include only consistently headed languages (both levels branch in the same direction). This models the idea that consistently headed languages are easier to process (Hawkins, 1994; Dryer, 1992; Haider, 1997), and so are more likely to be learned¹.

We can run a chain of iterated learning and record the hypothesis adopted by the learner at each generation. By doing this for many independent chains, we can calculate the proportion of learners that adopt a certain hypothesis (or set of hypotheses) at a given generation (i.e. the probability of a learner adopting a given hypothesis at a given generation). This is what I will refer to as the distribution over hypotheses².

6.2.1.1 Hypothesis space

The hypotheses in the model include any (un-ordered) combination of the above four languages. There are 15 possible hypotheses: 4 monolingual (e.g. $[\{Left, Left\}]$), 6 bilingual (e.g. $[\{Left, Left\}, \{Right, Right\}]$), 4 trilingual and one quadrilingual ($[\{Left, Left\}, \{Right, Right\}, \{Left, Right\}, \{Right, Left\}]$). See figure 6.2 for a graphic representation. Note that while the original monolingual models had a symmetrical hypothesis spaces (see Ferdinand and Zuidema, 2009, p. 1790), the hypothesis space here is not symmetrical. Being bilingual does not incur a penalty of any kind. In this model, the first agent is initialised with a random monolingual hypothesis.

6.2.1.2 Likelihood

The probability selecting a language from hypothesis h_i that can produce sentence d is the proportion of languages in hypothesis h_i that can generate each variant in d :

$$p_s(d_i|h_i) = \frac{|d \in h_i|}{|h_i|} \quad (6.1)$$

¹As mentioned in the previous chapter, there are many ways of interpreting the prior, for instance the amount of data that a learner needs to see before rejecting a certain hypothesis.

²Note that this is not necessarily the stable distribution, which is the probability of an agent adopting a given hypothesis over an infinitely long chain. However, 10 generations appears to be enough for this model to converge on a stable distribution: the same analysis run with 100 generations did not produce significantly different results, and the stable distribution calculated using the first eigenvector of the transition matrix for 100,000 runs for each hypothesis also returned equivalent results.

The perception of each variant may be affected by noise. A variant may be perceived as any variant with probability $\frac{n}{v}$ where v is the number of possible variants in the system ($v=2$ in the current example). Therefore, when a learner perceives data d , the likelihood of it being produced by hypothesis h_i is

$$p(d|h_i) = \prod_{i=1}^{|d|} \begin{cases} p_s(d_i|h_i)(1 - \frac{n}{v}) & \text{if } d_i \in h_i \\ \frac{n}{v} & \text{if } d_i \notin h_i \end{cases} \quad (6.2)$$

6.2.1.3 Prior

The prior is calculated in the following way. There is a set of hypotheses H containing hypotheses $h_1, h_2, h_3...$ and a sub-set of hypotheses H_b that are easier to process. There is a prior parameter α ranging from 0 to 1. α is the proportion of the prior that is shared by the favoured hypotheses (here, the consistently-headed hypotheses):

$$p(h_i) = \begin{cases} \frac{1}{|H|^{\frac{1}{1-\alpha}}} \frac{\alpha}{1-\alpha} & \text{if } h_i \in H_b \\ \frac{1}{|H|^{\frac{1}{1-\alpha}}} & \text{if } h_i \notin H_b \end{cases} \quad (6.3)$$

H_b in this model is defined as any hypothesis containing only consistent languages. In the current example there are 3 hypotheses in H_b : $[\{Left, Left\}]$, $[\{Right, Right\}]$ and $[\{Left, Left\}, \{Right, Right\}]$. This leaves 12 hypotheses that have a lower prior bias. In a model where $\alpha = 0.6$, the prior bias of $[\{Left, Left\}]$ (consistent) is 0.09, and the prior bias of $[\{Right, Left\}]$ (inconsistent) is 0.06.

6.2.1.4 Posterior

The posterior probability of hypothesis h_i given data d is calculated using Bayes' law:

$$p(h_i|d) = \frac{p(d|h_i)p(h_i)}{p(d)} \quad (6.4)$$

Since $p(d)$ is constant, the posterior is proportional to the numerator of this equation. A learner receives data and evaluates the posterior probability of each hypothesis in the hypothesis space. The learner then selects a hypothesis to generate variants for the next generation. The hypothesis is either chosen by sampling the hypotheses with probabilities proportional to the posterior probability (sampling), or selecting the single hypothesis with the highest posterior probability (maximum a posteriori, see section 5.4.4 in the last chapter)

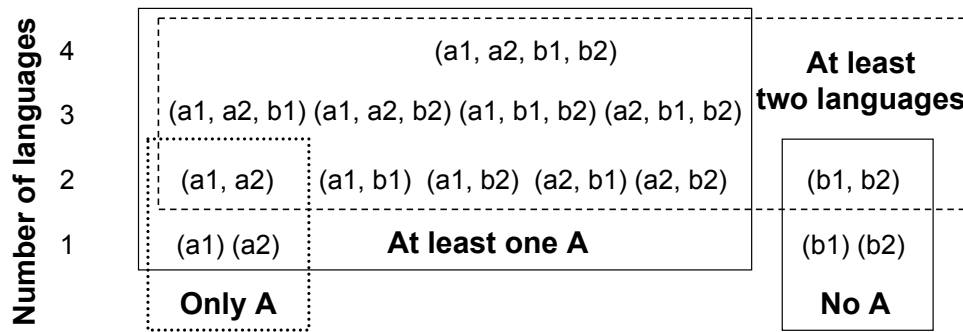


Figure 6.2: A representation of the space of hypotheses. There are 4 languages of two types (a1, a2, b1 and b2) representing consistently headed (a) and inconsistently headed (b) languages. Languages within brackets belong to the same hypothesis. Different criteria can be used to break up the space into different partitions.

6.2.2 Analysing the distributions

When considering a Bayesian model where learners adopt a single variant, the analysis of the dynamics were straightforward because there was an obvious unit of measurement: the distribution of variants. However, allowing individuals to adopt multiple hypotheses creates more possible analyses because there are more kinds of ways of measuring the distribution of hypotheses. We could measure the proportion of consistent hypotheses in the population of hypotheses. In the ‘monolingual’ model, this is the same as counting the proportion of consistent hypotheses in the population of learners. However, by allowing the adoption of multiple hypotheses, the population of hypotheses and the population of speakers are now not so straightforwardly linked. We could measure the proportion of agents speaking only consistent languages, or the proportion of agents speaking at least one consistent language. Finally, we could look at how the number of hypotheses are distributed over speakers, for example counting the proportion of agents who speak at least two languages. This is an approximation of ‘bilingualism’ in the sense that individuals ‘speak’ multiple ‘languages’. However, this is not bilingualism as defined in this thesis, simply because there is no social variable on which to condition the ‘languages’. Figure 6.2 shows a representation of the space of hypotheses with different partitions. Note that we could go further to look at the proportions of speakers speaking 1,2,3 and 4 languages separately.

6.2.3 Results for single languages

The following sections present the results of the first model³. We’ll start by looking at a simple case. The hypothesis space can be limited to hypotheses with only one language. This is a ‘monolingual’ model. This will replicate the results

³The implementation of this model was built on top of code made available by Kirby and Smith (2010)

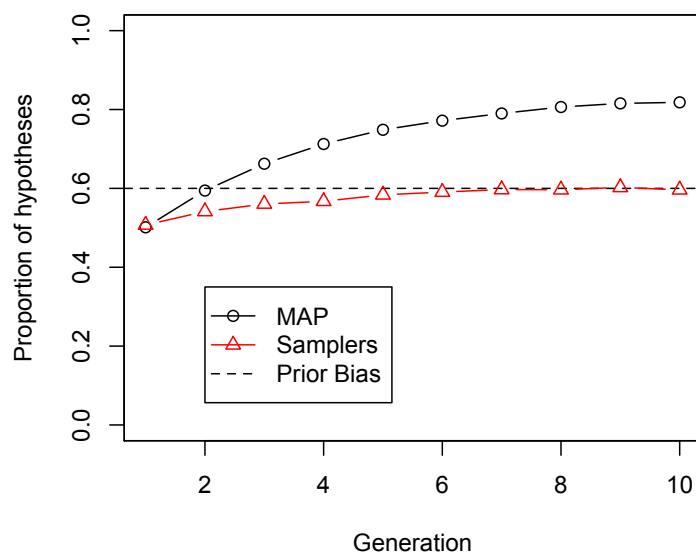


Figure 6.3: Proportion of consistently headed languages in a chain of Monolingual Samplers (red squares) and MAP learners (blue diamonds). Generations are initialised with equal amounts of data from each type of hypothesis.

of models that did not allow learners to adopt multiple languages. Figure 6.3 shows the average results for 10,000 runs of the model with a bias of 0.6 and a bottleneck of 5 utterances. For samplers, the proportion of consistently headed languages converges on the prior bias of the learners (60%), as in Griffiths and Kalish (2007). For MAP learners, this bias is exaggerated (non-convergence), as in Kirby et al. (2007).

6.2.4 Results for multiple languages

We can now look at the model dynamics when agents can adopt multiple ‘languages’. Figure 6.4 shows the results for the sampling learners, including the different measures mentioned above. Again, these are the averages for 10,000 runs with a bias of 0.6 and a bottleneck of 5 utterances. The proportion of consistent languages in the population of languages (a single agent may have more than one language) is lower than in the monolingual case, just above the 50% level. The proportion of agents speaking only consistent languages is much lower and the number of agents speaking at least one consistent language is much higher. This means, of course, that the proportion of bilingual agents increases.

Figure 6.5 shows the same experiment with MAP learners. Here, the proportion of bilingualism is much lower, while the proportion of agents speaking only consistent languages is higher and the proportion of agents speaking at least one consistent language is higher than in the monolingual case, but equal to the sam-

pling chain. The main difference between the samplers and MAP learners, then, is the prevalence of bilingualism that tends to emerge.

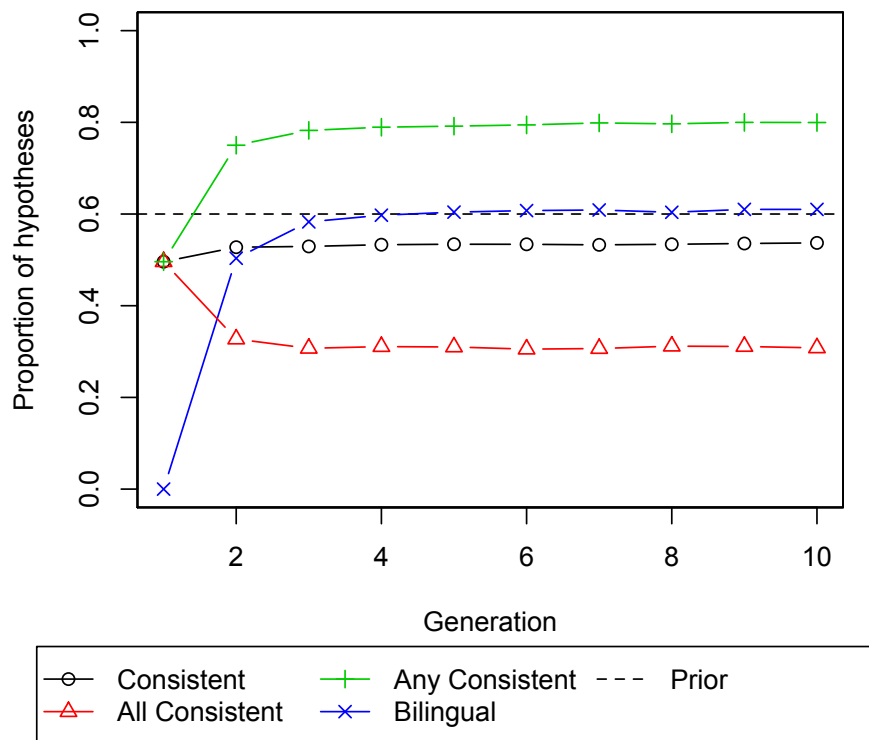


Figure 6.4: Results for chains of bilingual samplers. Proportion of consistent languages in population of languages (black circles); agents speaking more than one language (red triangles); agents speaking only consistent languages (green crosses); agents speaking at least one consistent language (blue xs)

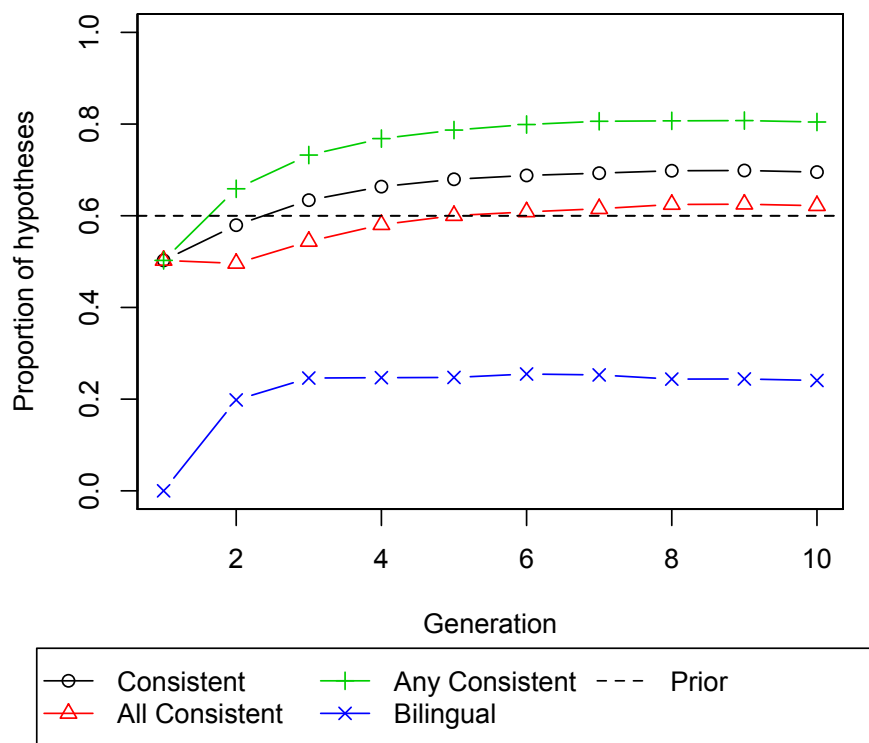


Figure 6.5: Results for chains of bilingual MAP learners. As in figure 6.4.

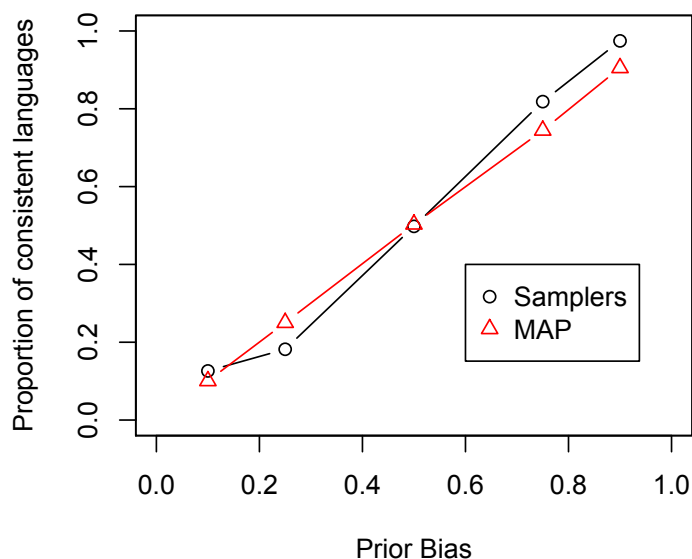


Figure 6.6: Average distributions of consistently headed languages after 10 generations for 10,000 chains of Monolingual MAP learners (red triangles) and Samplers (black circles) for varying strengths of bias.

6.2.5 Effects of the bias

Figure 6.6 shows how the proportion of consistently headed languages changes with the bias strength for learners who only adopt one language⁴. For Samplers, the relationship is perfectly linear. That is, the stable distribution of languages will always reflect the prior. The MAP learners tend to emphasise the bias and the relationship is an S-curve. This result is in line with previous findings (Griffiths and Kalish, 2007; Kirby et al., 2007).

Figures 6.7 and 6.8 show how other measures of the hypotheses' distribution behave under different bias strengths in a model that allows learners to adopt multiple hypotheses⁵. The relationship with the number of consistently headed languages is not linear any more. There is not a symmetrical distribution around 0.5 because there are an uneven number of hypotheses that include only consistently headed languages (3 hypotheses out of 15). In the 'monolingual' model, there were an even number of hypotheses that received the maximum prior bias as hypotheses that received the minimum prior bias. Note, however, that defining

⁴Calculated from the average final distributions of 10 generations for 10,000 chains of Monolingual MAP learners.

⁵Calculated from the first eigen vector of the transition matrix. To construct the transition matrix 100,000 chains of one generation were run initialised with each hypothesis (total 1.5 million generations). Where necessary, Laplace smoothing was used in order to ensure ergodicity in the transition matrix with MAP learners.

a rule over languages in a combinatoric system is not straightforward (consider figure 6.2)⁶.

There are more bilinguals with a lower bias since, as the bias increases, it tends to make up a larger part of the posterior probability and ‘over-rides’ the likelihood taken from the data.

For MAP learners, for higher biases, the distribution of consistent languages behaves like a chain of single individuals adopting single hypotheses (compare blue lines in figure 6.6 and 6.8). But for lower biases, the proportion of agents with bilingual hypotheses in the stable distribution is greater than the prior. Individuals with a high prior are likely to converge on the hypothesis with only consistent languages. There are mostly single-language hypotheses, with a small proportion being ‘bilingual’. Indeed, the probability of ‘bilingual’ hypotheses is proportional to the inverse of the probability of consistent hypotheses.

Figure 6.9 shows the difference in the stable distributions between models run with MAP learners and samplers. As expected, a larger proportion of the stable distribution is taken up with bilingual hypotheses in a chain of sample learners than in a chain of MAP learners. This difference is greatest with moderately strong biases (between 0.6 and 0.9). However, there is little difference between the two types of learner for the stable distribution of hypotheses containing at least one consistent language.

Previous models have discussed whether the distribution of hypotheses converges on the prior probability (see the last chapter). Allowing a form of bilingualism in this model forces us to choose the kind of bias that applies and the kind of distribution we measure. Some measures differ greatly from the prior. For example, the proportion of bilingual hypotheses is inversely and non-linearly related to the prior. This isn’t surprising because the prior is not explicitly defined over the number of languages in a hypothesis.

The distribution that most closely reflects the prior bias is the proportion of hypotheses with only consistent languages. However, this distribution does not converge to the prior either and is also affected by the space of possible language combinations. This demonstrates that the question of whether the learning bias is reflected in observable distributions depends on our assumptions about how the prior is related to the unit of the distribution and the space of hypotheses.

⁶In fact, there are some systems of combinations of X in Y numbers that are impossible to partition into sub-sets of equal size because they have an odd number of combinations. A general solution for determining the minimum number of rules for dividing categorical features into even subsets would be related to the partition problem (Mertens, 2006) or the bin-packing problem (Lee and Lee, 1985; Martello and Toth, 1990), and would be likely to be NP-hard.

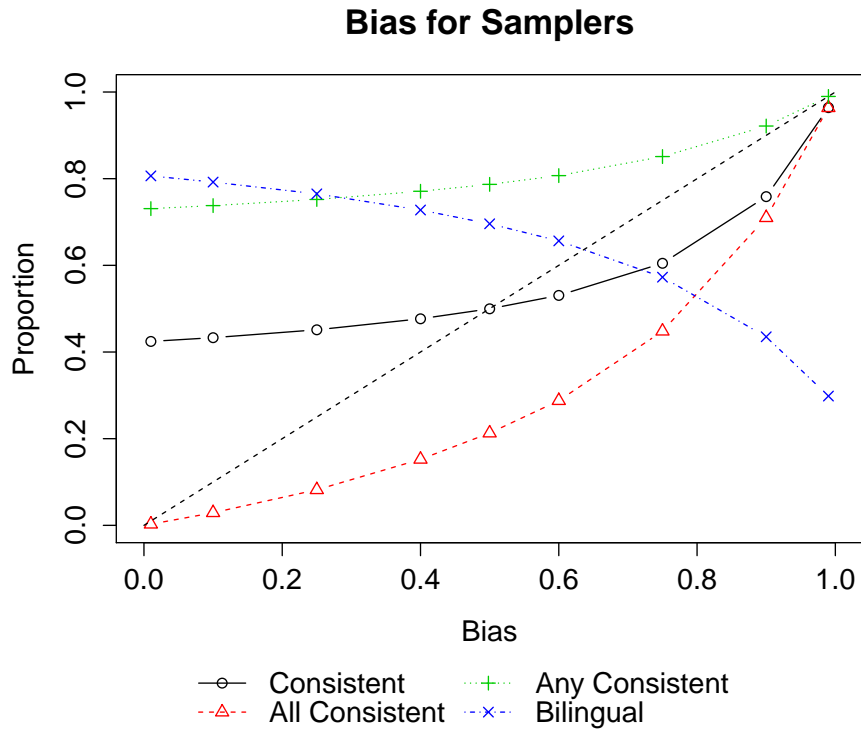


Figure 6.7: Stable distributions for ‘bilingual’ samplers with varying biases. Proportion of consistent languages in population of languages (black circles); Proportion of agents speaking only consistent languages (red triangles); Proportion of agents speaking at least one consistent language (green crosses). Proportion of agents speaking more than one language (blue xs); The dashed horizontal line indicates the distribution equal to the prior bias. Noise = 0.2.

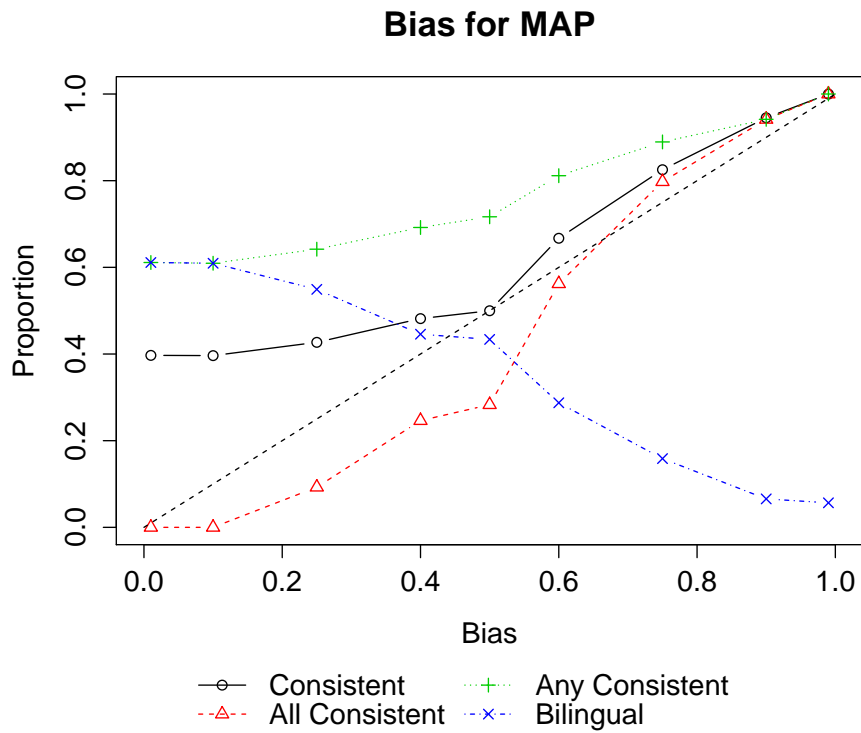


Figure 6.8: Stable distributions for ‘bilingual’ MAP learners with varying biases. Proportion of consistent languages in population of languages (black circles); Proportion of agents speaking only consistent languages (red triangles); Proportion of agents speaking at least one consistent language (green crosses). Proportion of agents speaking more than one language (blue xs); The dashed horizontal line indicates the distribution equal to the prior bias. Noise = 0.2.

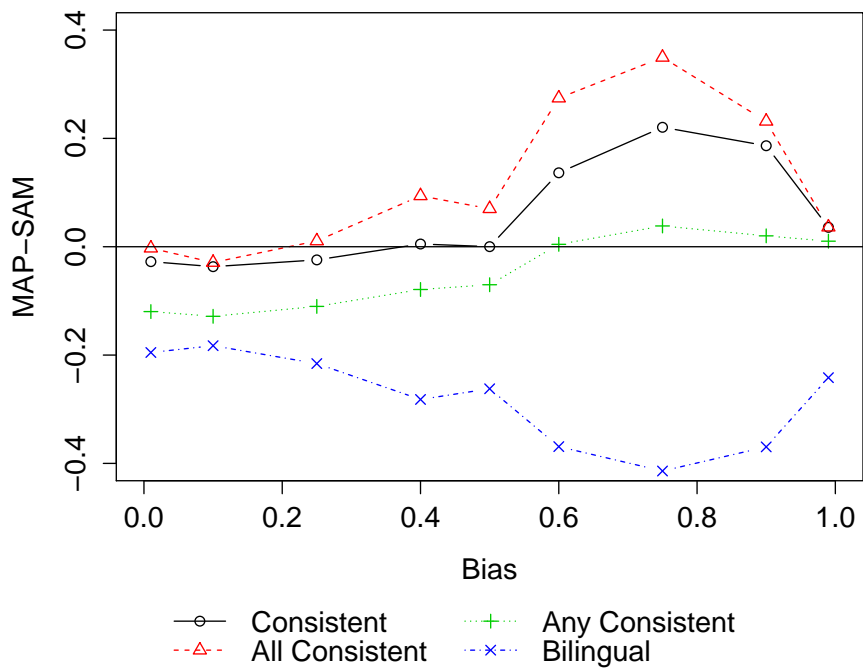


Figure 6.9: The difference in stable distributions between ‘bilingual’ MAP learners and samplers as a function of the prior bias. A positive value indicates that the proportion of a given hypothesis type in the stable distribution is higher for the MAP learners than the samplers. Data and legend is the same as in figures 6.7 and 6.8.

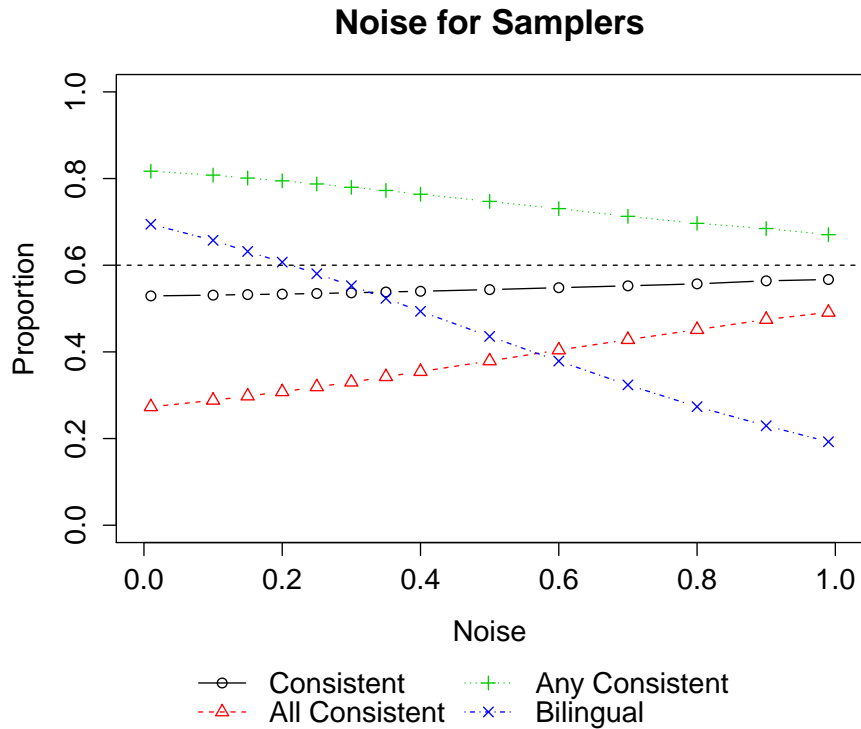


Figure 6.10: Stable distributions for chains of bilingual Samplers with varying amounts of noise. $\alpha = 0.6$.

6.2.6 Noise

There are also small interactions between the distributions of languages and the amount of noise interfering with the transmission of sentences from teacher to learner. Figures 6.11 and 6.10 show how the distributions change as the level of noise increases⁷. Note that with a noise level of zero (and with enough observations), the learners will never choose a different hypothesis from their teacher. For the MAP learners, the peaks and troughs in the middle of the distribution are due to interactions with the size of the bottleneck. The results for samplers is clearer, with the number of bilinguals falling as the level of noise increases, due to more agents speaking only consistently headed languages. Bilingualism (as well as the chain dynamics in general) is more stable under noise for samplers than MAP learners. Figure 6.12 shows the difference between the MAP and sampler results.

⁷Calculated from the first eigen vector of the transition matrix. To construct the transition matrix 100,000 chains of one generation were run initialised with each hypothesis (total 1.5 million generations). Where necessary, Laplace smoothing was used in order to ensure ergodicity in the transition matrix with MAP learners.

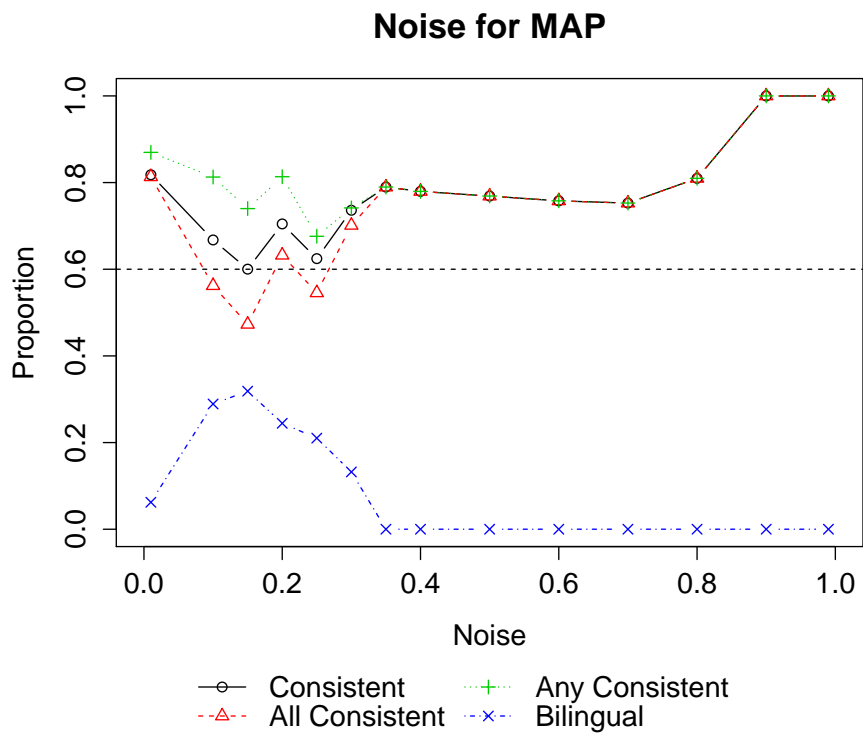


Figure 6.11: Stable distributions for chains of bilingual MAP learners with varying amounts of noise. $\alpha = 0.6$.

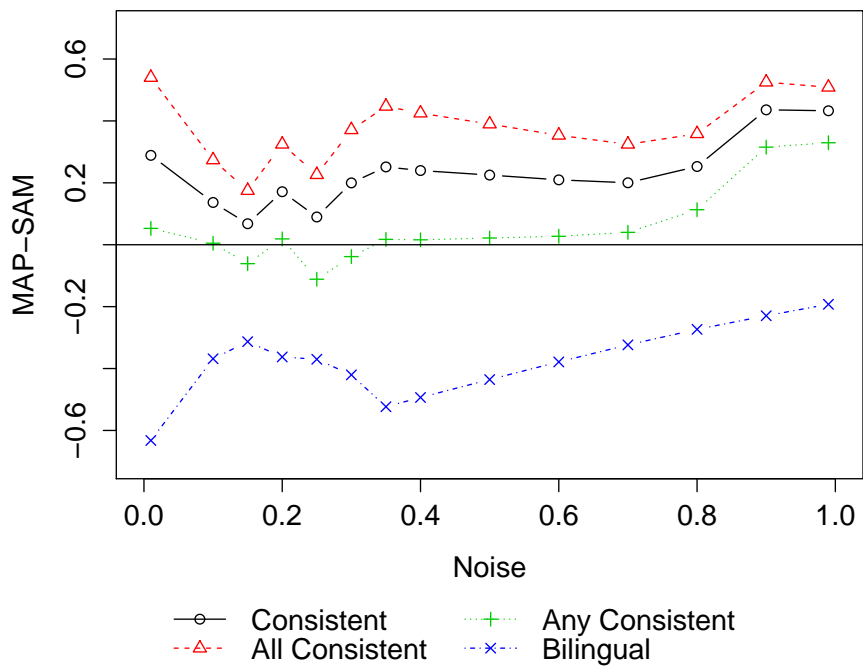


Figure 6.12: The difference in stable distributions between ‘bilingual’ MAP learners and samplers as a function of the noise level. A positive value indicates that the proportion of a given hypothesis type in the stable distribution is higher for the MAP learners than the samplers. Data and legend is the same as in figures 6.10 and 6.11.

6.2.7 Conclusions of model 1

The model above adapted a model of Bayesian iterated learning and allowed agents to adopt more than one language. As described above, this model violates the rationality assumption, by not fully taking into account how individuals select hypotheses. More complex models are available which address this (see previous chapter and Burkett and Griffiths, 2010; Smith and Thompson, 2012). However, this model has highlighted two points. Firstly, allowing individuals to adopt multiple ‘languages’ allows many different kinds of analysis. Specifically, by introducing the possibility of bilingualism, a distinction arises between the population of languages and the population of agents. In models with a chain of single individuals, this distinction was not relevant. One of the main debates discussed in the last chapter is whether there was a straightforward link between the biases of learners and the distribution of phenomena in the world. This model demonstrates that the answer to this question depends on the kind of unit that the biases and distributions are based on. Moreover, it depends whether you assume that there is a real cognitive bias over the unit that is being measured. For instance, this thesis argues that the unit of a language is not directly observable to children (see chapter 3), so it might not be valid to assume a cognitive prior over the number of languages to expect (see section 5.9.1 in chapter 5). In this case, the question of how a prior bias over low-level phenomena is linked to a higher-level, emergent phenomena may be a qualitatively different question than those addressed by previous models.

Secondly, the proportion of bilinguals that emerge in a Bayesian Iterated Language framework depends on the learning algorithm involved (sampling vs. MAP). While the choice of learning algorithm largely depends on the modeller’s assumptions, it is possible to see the two mechanisms as points on a continuum (Kirby et al., 2007; Blythe, 2009). If agents are able to choose between different learning algorithms (or choose an intermediate point between the two) to fit the input, then different kinds of input may sustain different proportions of ‘bilingualism’, as defined in this model. However, as the bottom-up model presented in the next chapter will show, drawing conclusions about bilingualism from this kind of top-down model can be invalid.

The next section extends the model again to include multiple speakers. Models with multiple speakers and multiple hypotheses have already been investigated (e.g. Smith, 2009; Burkett and Griffiths, 2010; Kirby et al., 2012). However, while other models introduced some assumptions to maintain full rationality, the assumptions in the models in this chapter are made in order to investigate bilingualism. Different conclusions come out of these approaches, demonstrating that the research questions of researchers can affect the assumptions that go into a model, but also that the constraints of the model can affect the kinds of research questions being asked. Furthermore, these conclusions will contrast again with those of the bottom-up model developed in chapter 7.

6.3 Model 2: Multiple learners and multiple languages

The previous model had individuals with biases over particular languages. The next model is slightly different in that individuals have biases over the amount of variation to expect in a teacher's utterances. Rather than adapting the model above, I define the next model in its entirety.

6.3.1 Model definition

Teachers know a set of languages (their adopted hypothesis). Teachers produce utterances for learners to observe. They do this by selecting a language at random from their hypothesis and producing an utterance from that language. However, the observation may be affected by noise so that an utterance can be perceived as belonging to another language. Learners must induce the most likely hypothesis that their teacher is using to produce these utterances. They do this through Bayesian induction: They calculate the likelihood of each hypothesis producing the data they observe and combine it with a prior probability for that hypothesis to give a posterior probability. Learners then adopt the hypothesis with the highest posterior probability (maximum a priori learners).

6.3.1.1 Hypothesis space

There are m possible utterances which belong to m languages. For example, two possible utterances a and b , belonging to two possible languages A and B respectively. Teachers can know a sub-set of all possible languages. We call this sub-set the hypothesis. In the example with two languages, there are three hypotheses: A , B and AB . So, utterances belong to languages, and languages belong to hypotheses. Some hypotheses have languages in common.

6.3.1.2 Likelihood

The likelihood of data being produced by a hypothesis is the following:

$$p(d|h) = \prod_{i=1}^{|d|} \begin{cases} \frac{1}{|h|}(1-n) + \frac{1}{|h|}n\frac{|h|}{m}, & \text{if } d_i \in h \\ \frac{1}{|h|}\frac{|h|n}{m} & \text{if } d_i \notin h \end{cases} \quad (6.5)$$

An utterance may be perceived to belong to a hypothesis in two ways: First, the teacher produces the utterance from a language in the hypothesis (with probability $\frac{1}{|h|}$), and it is unaffected by noise (probability $1-n$). Alternatively, the teacher's utterance may be affected by noise (probability n), but is still perceived to belong to a language within the hypothesis (the probability being the size of the hypothesis divided by the number of possible languages $\frac{|h|}{m}$). The rational approach to this is to assume that, if the utterance is part of the hypothesis that is being evaluated, then the likelihood of it being produced by that hypothesis

is the probability of a teacher selecting the corresponding language ($\frac{1}{|h|}$) times the probability of it being unaffected by noise ($1 - n$), plus the probability of a teacher selecting the corresponding language ($\frac{1}{|h|}$), it being affected by noise (n) but being changed into an utterance that is still part of the hypothesis.

If an utterance is perceived not to belong to a language in the hypothesis, then this could still have been produced by hypothesis h if it was affected by noise (probability n) and ‘changed’ to an utterance belonging to a language not in the hypothesis (with probability $\frac{|h|}{m}$).

6.3.1.3 Prior

The agents in this model have a prior expectation of the number of languages that a teacher will speak. The prior probability is based on a beta function. This is a function that can describe several types of relationships between the number of languages in a hypothesis and the prior probability the learner assigns to that hypothesis. The relationship is affected by two parameters. Figure 6.13 shows how this relationship changes with different settings of the two parameters. For instance, the ‘monolingual’ bias is represented by the blue line and assigns a high probability to hypotheses where individuals speak fewer languages. The ‘bilingual’ bias represented by the red line is the exact opposite⁸. For a model with m possible languages the beta value of hypothesis h is:

$$beta(h) = F\left(\frac{|h|}{m+1}; \alpha, \beta\right) \quad (6.6)$$

where F is the beta distribution function with parameters α and β and $|h|$ indicates the number of languages in hypothesis h . This value is divided by the number of possible languages *plus one* because the beta distribution function only accepts values less than 1. The prior probability of a hypothesis in a set of hypotheses H is then calculated as the beta value divided by the the sum of beta values for all hypotheses.

$$p(h) = \frac{beta(h)}{\sum_{i=0}^{|H|} beta(h_i)} \quad (6.7)$$

⁸While this is a bias that fits the research question, it is less clear how this abstract bias would actually relate to a real cognitive bias. Minimally it requires that learners have a mechanism that can differentiate between languages. Infants appear to have statistical learning abilities that can do this (see section 3.9 on page 41). It is less clear what units this bias would operate on. Since linguistic systems vary widely in the amount of variation they exhibit at phonetic, lexical and syntactic levels, it is difficult to imagine such a bias being effective considering only a single domain. However, the levels of complexity in each domain may trade-off against each other (Winters, 2011), meaning that a general measure of variation could be expected. An alternative solution is to use general learning biases, as suggested in the bottom up model presented in chapter 7.

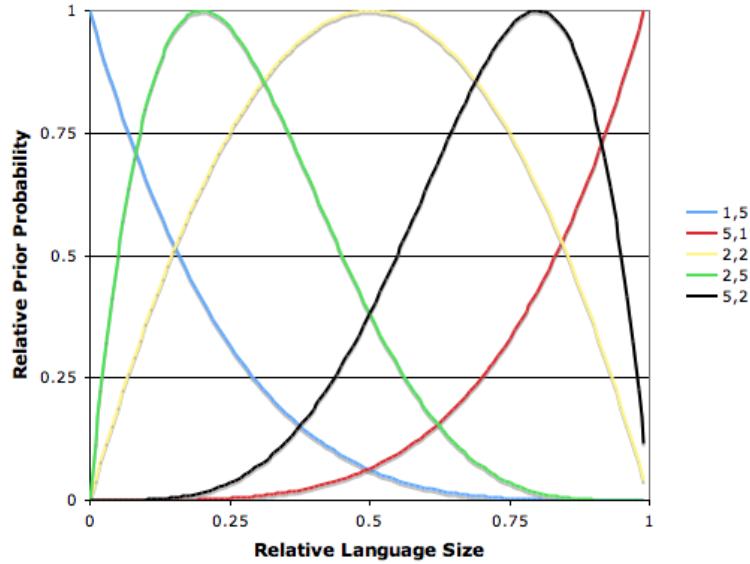


Figure 6.13: Beta Distributions for different values of α (1st number in legend) and β (2nd number in legend).

This ensures that the total prior probabilities for all hypotheses sum to 1.

6.3.1.4 Posterior

The posterior probability for a hypothesis follows Bayes' law by combining the likelihood with a prior probability $p(h_i)$:

$$p(h_i|d) = \frac{p(d|h_i)p(h_i)}{\sum_{i=1}^h p(d|h_i)p(h_i)} \quad (6.8)$$

The term at the bottom normalises the probabilities so that they sum to one.

6.3.2 An example

Below is a simple example with two languages A and B , which yields 3 hypotheses A , B and AB . The noise probability n is set to 0.1. For now, let's assume a prior bias that is constant for all hypotheses (so the posterior probability is directly proportional to the likelihood). We'll also assume that the learner only hears one utterance. This means there are two possible situations - either the learner hears a or the learner hears b . If a learner observes a , then the likelihood of each hypothesis ($p(d|h_i)$) is as follows:

$$\begin{aligned} p(a|A) &= \frac{1}{|h|}(1-n) + \frac{1}{|h|}n\frac{|h|}{m} = \frac{1}{1}(1-0.1) + \frac{1}{1}0.1\frac{1}{2} = 0.95 \\ p(a|B) &= \frac{1}{|h|}\frac{|h|n}{m} = \frac{1}{1}\frac{1 \times 0.1}{2} = 0.05 \\ p(a|AB) &= \frac{1}{|h|}(1-n) + \frac{1}{|h|}n\frac{|h|}{m} = \frac{1}{2}(1-0.1) + \frac{1}{2}0.1\frac{2}{2} = 0.5 \end{aligned} \quad (6.9)$$

Since the prior bias is constant for any hypothesis ($p(h_i) = \frac{1}{3}$), then the posterior can be calculated as:

$$\begin{aligned}
 p(A|a) &= \frac{p(d|h_i)p(h_i)}{\sum_{i=1}^h p(d|h_i)p(h_i)} = \frac{0.95 \times \frac{1}{3}}{0.5} = 0.6\dot{3} \\
 p(B|a) &= \frac{p(d|h_i)p(h_i)}{\sum_{i=1}^h p(d|h_i)p(h_i)} = \frac{0.05 \times \frac{1}{3}}{0.5} = 0.0\dot{3} \\
 p(AB|a) &= \frac{p(d|h_i)p(h_i)}{\sum_{i=1}^h p(d|h_i)p(h_i)} = \frac{0.5 \times \frac{1}{3}}{0.5} = 0.\dot{3}
 \end{aligned}
 \tag{6.10}$$

As expected, the posterior probability of hypothesis A given that the learner hears a is higher than the other two hypotheses. The same process can be followed to calculate the posterior probability of each hypothesis if the learner observes a b :

$$\begin{aligned}
 p(A|b) &= 0.0\dot{3} \\
 p(B|b) &= 0.6\dot{3} \\
 p(AB|b) &= 0.\dot{3}
 \end{aligned}
 \tag{6.11}$$

We can now work out the probabilities of transitions between hypotheses from a teacher to a learner, if we assume single teachers and single learners. For example, the probability of a transition from hypothesis A to hypothesis B is the posterior probability $p(A|a)$ times the probability of A producing a , $p(a|A)$, plus the posterior probability $p(A|b)$ times the probability of A producing b . Figure 6.14 shows these transition probabilities.

The interesting thing to note is that the transitions in and out of AB are

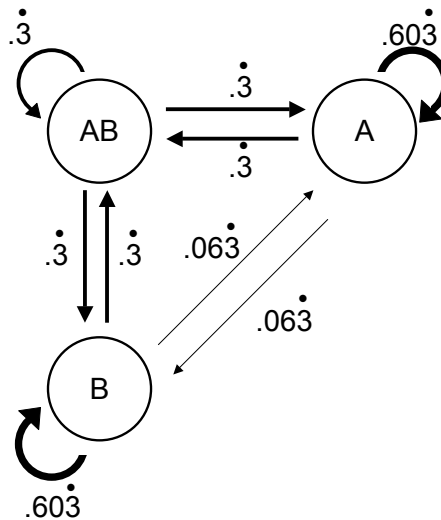


Figure 6.14: The probabilities of changing hypothesis state between a teacher and a learner in a system with two languages and a uniform prior bias.

perfectly balanced, while there is an asymmetry in the transitions for A and B .⁹ The ‘monolingual’ hypotheses A and B are more ‘learnable’ than the ‘bilingual’ hypothesis AB , in the sense that a learner learning from a teacher who knows A is more likely to adopt A than any other hypothesis. We might intuitively expect individuals learning in this system to be monolingual more often than bilingual.

It is possible to calculate the stable distributions over hypotheses by taking the first eigenvector of the transition matrix. The stable distribution is the probability that a given learner in an infinitely long chain of teachers and learners will have a certain hypothesis. It turns out that the stable distribution is shared equally across all hypotheses with a uniform prior bias (i.e. convergence to the prior). That is, observing a bilingual speaker is as likely as observing a monolingual speaker of either language. Even though the ‘monolingual’ hypotheses A and B are more ‘learnable’, because the hypothesis AB sits in a transition phase between them and so the population-level distribution is skewed¹⁰.

This result holds for more than two languages. However, while the stable distribution is constant over hypotheses, increasing the number of languages greatly increases the number of bilingual hypotheses. This means that with 3 or more

⁹We can show that, in general, the probability of a transition from the hypothesis that contains all the language types (h_{omni}) to itself is constant and independent from noise. Given the set C of all possible combinations of k utterances of m languages, the probability of transitioning from a given hypothesis to h_{omni} is

$$\sum_{i=1}^{|C|} \left(\frac{1-d}{|h_B|} + \frac{d}{m} \right)^{|C_i \cap h_B|} \left(\frac{d}{m} \right)^{|C_i \setminus h_B|} \quad (6.12)$$

The probability of transition between h_{omni} and h_{omni} (the above equation where $h_B = m$) is

$$\sum_{i=1}^{|C|} \left(\frac{1-d}{m} + \frac{d}{m} \right)^{|C|} \quad (6.13)$$

$$= \sum_{i=1}^{|C|} \left(\frac{1}{m} \right)^{|C|} \quad (6.14)$$

This has a counter-intuitive result. It is intuitive that that any given utterance is more likely to be produced by the hypothesis that most closely and efficiently describes it (e.g. the utterance a is more likely to be generated by hypothesis A than any other hypothesis). However, the *average* likelihood for utterances generated by a teacher who knows all languages, for any hypothesis is equal. Put another way: a teacher who knows all languages generates many sets of utterances. The likelihood of each utterance set being generated by each hypothesis is calculated. The mean likelihood over sets of utterances will be the same for each hypothesis. This is the case for the teacher who knows all languages because they produce utterances from all languages with equal probability.

¹⁰The asymmetry between the transition likelihoods of monolingual and bilingual hypotheses leads to another point. The prior biases of learners may have a greater cumulative affect on the hypothesis they select when they receive data from a teacher who knows all languages than one who only knows one. This is a similar result to that of Burkett and Griffiths (2010) who show that the stable distributions are sensitive to the amount of variation that learners expect. However, a formal proof linking these effects is beyond the scope of this chapter.

languages, individuals in this system are more likely to be bilingual than monolingual.

6.3.3 Chain dynamics

We can now see how the prior bias affects the stable distribution. First, we assume only one teacher per generation. The distribution over languages over many generations can be calculated, given different prior biases. The system considered will include 3 languages and the number of utterances a learner hears is 5.

Three types of bias are used: a monolingual bias which exponentially favours monolingual hypotheses (beta distribution parameters $\alpha = 1, \beta = 10$), a uniform bias that favours all hypotheses equally ($\alpha = 1, \beta = 1$) and a bilingual hypothesis that exponentially favours hypotheses with many languages ($\alpha = 10, \beta = 1$). The hypothesis space contains three languages. The noise level is set to 0.05.

Figure 6.15 shows the stationary distribution over hypotheses for these three biases¹¹. For the uniform bias, each hypothesis is equally likely, reflecting the prior bias. For the monolingual and bilingual biases, the distribution is skewed towards the three monolingual hypotheses and the bilingual hypothesis respectively. This reflects the prior bias.

However, note that for the bilingual bias, the prior bias is spread over a single hypothesis which contains all the languages, whereas for the monolingual hypothesis the prior is spread over *three* hypotheses. That is, the stability (how likely the learner is to adopt its teacher's hypothesis) of the chain with a monolingual bias is lower than that of the chain with the bilingual bias.

6.3.3.1 Memory constraints

Note also that, for the uniform bias, the amount of 'bilingualism' (hypotheses with more than one language) is greater than the amount of 'monolingualism' (hypotheses with only one language). This is due to there being more bilingual hypotheses than monolingual hypotheses.

A hypothesis space can be constructed where the number of monolingual and bilingual hypotheses are equal. For example, a hypothesis space with 3 languages, but with a maximum of two languages in a hypothesis. This yields the hypotheses A, B, C, AB, AC and CB . Figure 6.16 shows the stationary distribution of a chain with such parameters. Now, the monolingual and bilingual distributions are evenly spread over the same number of hypotheses (the uniform bias yields a uniform distribution).

¹¹Calculated by Monte Carlo sampling of the transition probabilities between each pair of hypotheses over 100,000 runs, then taking the first eigenvector of the transition matrix.

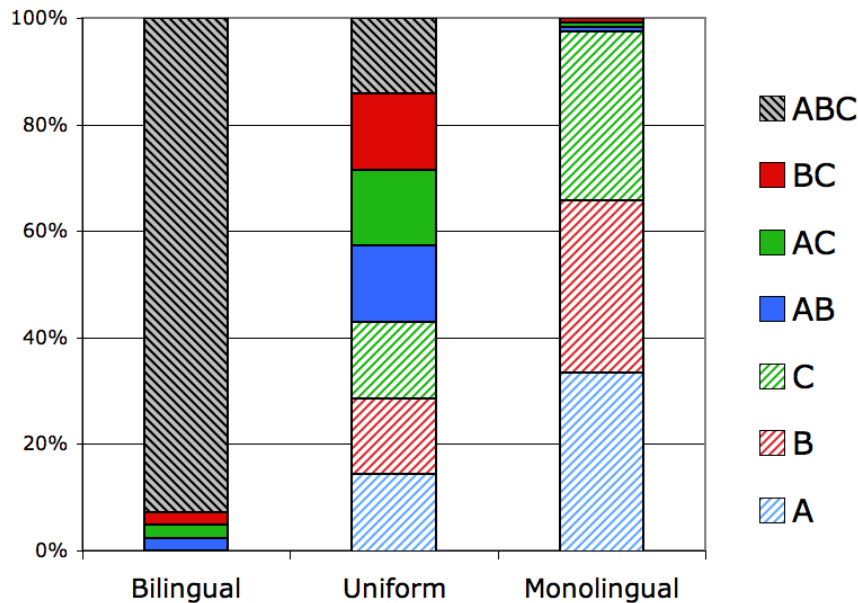


Figure 6.15: Stationary distributions over hypotheses for monadic chains of bilingual learners with the bilingual, uniform and monolingual biases. Monolingual hypotheses are indicated by coloured stripes, hypotheses with 2 languages are indicated by solid colours and the hypothesis with 3 languages is indicated by grey stripes.

For aspects of language learning that have few variants (e.g. aspects of syntax such as basic word order), then the number of bilingual possibilities is likely to be higher than monolingual possibilities because memory limits are likely to be large enough to accommodate many variants.

6.3.4 Allowing multiple teachers

We can extend the model to include multiple teachers. A learner is exposed to a *context*, which is a set of teachers. The learner then needs a function to choose a single hypothesis given the data they observe. The teachers all produce utterances, and the learner evaluates the most likely hypothesis for each teacher in the same way as above. Here we'll assume that the most likely context is this set of most likely hypotheses, and that the hypothesis that the learner adopts is the one that is a super-set of all languages in this context. That is, if a learner hypothesises that the most likely situation is that teacher 1 speaks language A and teacher 2 speaks languages A and B, then the learner will become a teacher who speaks languages A and B (the union of the languages in the context) ¹².

Given a population of teachers, a new generation is created in the following way. Teachers are paired randomly to form two-agent contexts. Each pair produces data for two new learners. These learners calculate the most likely context given

¹²This choice of process is slightly arbitrary, and is also not accounted for in the likelihood function, meaning that the individuals are not strictly rational.

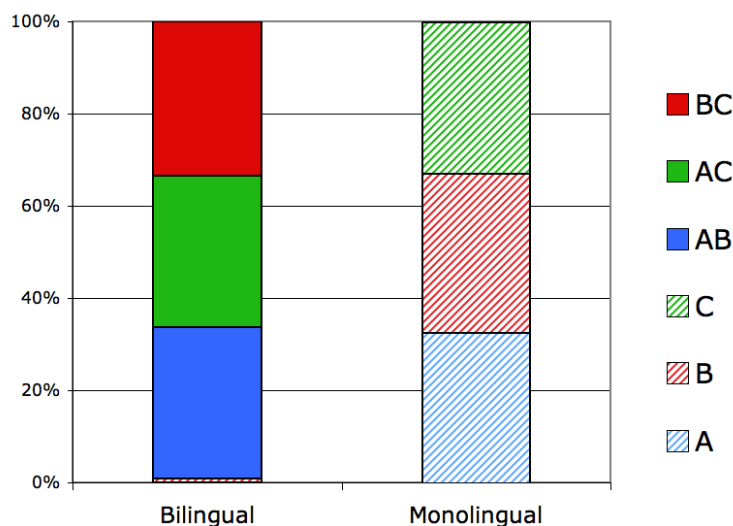


Figure 6.16: Stationary distributions over hypotheses with the bilingual and monolingual biases for a hypothesis space with an even number of monolingual and bilingual hypotheses. Distributions for some hypotheses are too small to be shown on the graph.

the data and their prior biases. The teachers are then replaced by the two learners, who become two new teachers as described above. The population is then paired randomly again and the process repeats. In this way, the number of agents in the population is kept constant.

We can look at how the prior bias affects population-level phenomena. For example, we can look at the distribution of languages in the population, the number of languages in a population and the proportion of learners who induce the context with which they were presented (the number of ‘correct guesses’ by learners).

The model was run with various settings with a population of 10 agents for 100 generations. The agents were initialised with the same monolingual hypothesis. The noise level was set to 0.05 and teachers produced 10 utterances each. There were a total of four possible languages. Figure 6.17 shows results of a single run of the model. The graphs show the proportion of correct inductions by learners, the distribution of the number of languages spoken by agents and the number of different hypotheses in the population for three settings of the prior bias. In the leftmost column, the bias was set to favour monolingual hypotheses (beta distribution parameters $\alpha = 1, \beta = 10$), the middle column the bias was set uniformly ($\alpha = 1, \beta = 1$) and the rightmost column was set to favour bilingual hypotheses ($\alpha = 10, \beta = 1$). The beta distribution meant that, for the monolingual settings, for example, a hypothesis with a single language received a prior bias of 0.09, while a hypothesis with four languages received a prior bias of 3.4×10^{-7} . These values are switched for the bilingual bias.

The results show that a population does not generally evolve to become mono-

lingual. This is due to noise in the system and the probability of paired agents speaking different languages. Even with a strong monolingual bias, only one agent spoke only one language in the last twenty generations. An average of 9.5% of agents at each generation spoke two, 53% spoke three and 37% spoke four (also in the last 20 generations). For the uniform bias, no agents spoke one language, an average of 0.5% spoke two languages, 6% spoke three and 98.5% spoke four. For the bilingual bias, all agents spoke 4 languages in the last 20 generations.

A big difference between the monolingual and bilingual biases is the proportion of correct inductions by learners. The monolingual bias yields much lower correct inductions than the uniform and bilingual bias. (the model was run with the monolingual bias for 1000 generations and the correct induction proportion does not increase for the monolingual bias). Looking at the number of hypotheses in the population, the reason behind this becomes clear. The bilingual population quickly converge on the same hypothesis - the one that contains all the languages. It is able to do this because there is only one hypothesis that is given the highest prior bias (the one with all 4 languages). The monolingual population, however, have four hypotheses with the joint highest prior probability¹³. The monolingual learner is biased towards believing that its parents speak one language each. However, because the model only considers the number of languages each teacher speaks *independently*, these languages can be different, making the learner bilingual when it becomes a teacher. The monolingual population fluctuates around many hypotheses while the bilingual population can converge on one.

Even with no difference between the prior biases for hypotheses (the uniform distribution), the population tends to become bilingual. The proportion of correct inductions with a monolingual bias was actually significantly worse than with a uniform bias (mean proportion of correct inductions for the last 20 generations for monolingual bias = 6.5%, for uniform bias = 29.5%, $t = -6.4$, $df = 38$, $p\text{-value} < 0.0001$). The bilingual bias performed better than the uniform bias (for bilingual = 72.5%, $t = -8.7$, $df = 38$, $p\text{-value} = < 0.0001$).

Figure 6.18 shows the stable distribution over contexts (calculated from a transition matrix with each transition having a sample size of one million generations) and reflects much the same picture. For the monolingual bias, there are a large number of contexts with a high proportion, and many of these have mutually exclusive languages. For the bilingual bias, however, nearly 40% of the distribution is taken up by a single context, with the next 20% being highly compatible.

The stationary distribution revealed that, for the monolingual bias, 23.1% of agents spoke one language, 65% spoke two and 11.4% spoke three. For the bilin-

¹³This is similar to the result of Ferdinand and Zuidema (2009) who find that an asymmetrical hypothesis space may mean that samplers (who may adopt a range of hypotheses) are more stable than maximisers (who adopt a single hypothesis). However, all the individuals here are samplers.

gual bias, 0.017% spoke one language, 5.54% spoke two and 94.4% spoke three languages.

6.3.4.1 Memory constraints

The model predicts a strange outcome - that having a bilingual bias will tend to evolve individuals that are more alike, even if they are multilingual. However, this may be due to the constraints from the data. If there were many more languages than an agent could remember, then the bilingual population would not have the advantage of having a single favoured hypothesis to converge on.

In order to test this, the model was run with the same parameters, but with 5 possible languages. However, the maximum number of languages in a hypothesis was set to 3. This meant that there were 5 hypotheses jointly preferred by the monolingual bias and 10 hypotheses jointly preferred by the bilingual bias. Furthermore, no hypothesis contained all languages. The conversion from learner to teacher was also changed so that if the most likely context contained more than 3 languages, random languages were omitted from the largest hypothesis in the context until there were only 3 languages in the set of languages in the context.

It should be noted that, up to this point, the distribution of languages evolves under directed mutation. However, constraints on memory introduce an aspect of *selection*. Figure 6.19 shows the results. The monolingual correct induction proportions have increased and the number of hypotheses in the bilingual population is higher. However, a one-way ANOVA revealed a significant effect of bias type on correct induction proportions in the last 20 generations ($F(2,57) = 128.7$, $p < 0.0001$). The bilingual bias still yields a higher average correct induction proportion than the monolingual bias (for monolingual bias = 13%, for bilingual bias = 79.5%, $t = 17.8$, $df = 38$, $p\text{-value} < 0.001$). The monolingual bias also performs worse than the uniform bias (for uniform bias = 57%, $t = 9.8$, $df = 38$, $p\text{-value} < 0.0001$).

This can be explained in the following way. Bilingually biased teachers are likely to have more overlapping hypotheses than monolingual teachers. That is, they are likely to speak more languages, and so will have more languages in common. This means that languages that the parents did not have in common were more likely to be pushed out due to memory constraints (because there were fewer of them in the context, and languages are removed at random from individual hypotheses). Monolingually biased teachers, however, were more likely to have mutually exclusive hypotheses, leading to higher variation in the learners. This also explains why all three biases have similar numbers of hypotheses in the population, but have different proportions of correct inductions.

For the monolingual bias, no agents spoke only one language in the last 20 generations, an average of 19% spoke two and 81% spoke three. For the uniform bias,

no agents spoke only one language, one agent spoke two languages and the rest spoke three. For the bilingual bias, all agents in the last 20 generations spoke 3 languages.

6.3.5 Error thresholds

The result that the stationary distribution favours bilingual hypotheses is linked to the concept of error thresholds in biology (Eigen, 1971). The hypotheses in the model are sequences of discrete features, like a genome: One can think of each hypothesis as a sequence of m bits indicating whether a particular language is known or not. This introduces a sequence assumption. Many other Bayesian models have a one dimensional hypothesis space where each hypothesis is independent from the rest. Kalish et al. (2007) manipulate the amount of overlap between generations, but hypotheses spaces are usually at least symmetrical (see Ferdinand and Zuidema, 2009, p. 1790). The hypotheses in this model are not independent from each other: Some hypotheses contain others (e.g. the hypothesis $\{A, B\}$ contains the languages of the hypothesis $\{A\}$). This has three consequences. First, the hypothesis space is asymmetrical. Ferdinand and Zuidema (2009) demonstrate that the stable distribution for samplers gets further from the prior bias as the hypothesis space becomes more asymmetrical. This is reflected in the results above. Secondly, there's the possibility of an error threshold and thirdly there's the possibility of changing the fitness landscape.

An error threshold is defined as follows: For a population whose fitness is determined by a sequence of discrete elements (e.g. a genome or bitstring), an error threshold is approached as the number of elements in the sequence grows. This is when mutation spreads the distribution of sequence types over the whole hypothesis space. All information (sequence settings that contribute to high fitness) is lost due to mutation scrambling the sequences. This leads to the distribution over sequences approaching a uniform distribution. This is related to the likelihood in the model approaching a uniform distribution as the teacher speaks more languages: The hypothesis A is 'closer' to AB than it is to BC in the sense that A is more likely to 'mutate' into AB than BC (a learner with a parent who knows A is more likely to estimate the most likely hypothesis to be A, B than B, C).

This feature leads to the second consequence: The 'fitness landscape' (in the model, this is the proportion of the stationary distribution represented by a hypothesis) can be affected by the structure of the mutation dynamic. Because some hypotheses are 'closer' to each other, certain transitions are more likely, affecting the distribution of hypotheses that emerge. Indeed, we see that bilingual hypotheses are more likely to emerge, precisely because they are 'closer' to hypotheses with fewer languages than those hypotheses are to each other.

6.3.6 Conclusions of model 2

Individuals in model 2 considered hypotheses that were sets of languages. Because of the way these hypotheses were related (some being sub-sets of others), the population-level distribution over hypotheses was not straightforwardly related to the biases of individuals.

With multiple teachers, bilingualism is inevitable in this model, even with very strong innate biases against it. There are two driving forces for bilingualism according to this model. First, noise in transmission allows the system to move between states. Secondly, there is a learning bottleneck in the sense that a learner is not exposed to data from all members of the population. Therefore, it is likely that two individuals in a population will have different input. This creates parents that speak different sets of languages, leading to a tendency for a learner in the next generation to speak more languages than either of its teachers. Thus, bilingualism becomes the norm.

This model might suggest that research into bilingualism should focus on examining the strength of an individual's bias for monolingualism and whether the social structure causes bottlenecks in transmission. Appendix B uses model 2 to attempt to estimate the bias of individuals with real data, with limited success. The results are suggestive of human learners having a weak bias for monolingualism.

A potential problem for the model is that a learner's hypothesis is the result of collapsing the context of its parent generation. If some contexts are more likely to occur given this collapsing (and possible memory constraints), then a perfectly rational learner should take this into account. In fact, the population dynamics show that this is the case: hypotheses with more languages within them are more likely to emerge in a range of situations. However, there are two responses for this: First, how much of the structure of the problem is it reasonable to expect an individual to consider (also asked by Ferdinand and Zuidema, 2009)? The emergent properties of a population may be beyond the cognitive grasp of a limited learner, or too complex to integrate through evolution. There is also the point discussed in chapter 3 that increasing the psychological reality of cultural transmission factors does not help the anti-nativist argument. However, the second response is that the purpose of this model was to demonstrate that taking into account the distribution of variation over speakers is a rational thing to do. Building in a rational response to an emergent property of a population would only re-enforce this point. Put another way, building in a response to the emergent properties of the model may result in a different stationary distribution, but would still recognise the importance of the distribution of variation over speakers and the advantage of bilingualism in structured populations.

6.4 Conclusion

This section has presented two top-down models of iterated learning with Bayesian agents who can learn more than one language. Introducing the possibility of ‘bilingualism’ made the model qualitatively more complicated. Firstly, the hypothesis space is more complex, which affects the transmission dynamics. Secondly, there are many more ways to measure the linguistic properties of the population, each of which differs in its relationship with the biases of individuals. Finally, there is a greater potential for complexity in the learning algorithm of the individuals. Introducing variation also introduces the possibility of a bias over the expected amount of variance, as well as over specific languages. Furthermore, there are many ways to approach this variation, such as considering the variation within individuals’ utterances (this model), or in the utterances of the whole population (see the last chapter). A researcher’s choices regarding the assumption of rationality can also affect the assumptions about the individual’s learning algorithm (see also Ferdinand and Zuidema, 2009 and last chapter).

The second model helps us to understand part of the bilingual paradox: We have an intuition that learning two languages is harder than learning one, but bilingualism seems to be prevalent in the world. The second model showed that even though hypotheses with single languages are more ‘learnable’, this doesn’t mean that the distribution of hypotheses that emerge will favour monolingualism. This result partly stems from the encoding of bilingualism in this model as knowing a set of languages.

However, it is worth considering whether the phenomenon modelled here is really bilingualism. Certainly individuals can adopt multiple cultural features, but can these really be seen as multiple languages? Chapter 3 suggested that it might be invalid to represent whole languages as monolithic, discrete, static entities when studying bilingualism in a cultural evolution framework. Furthermore, the definition of bilingualism that this thesis uses is the amount of linguistic optionality that is conditioned on social variables. There are a few problems with measuring this in the current model. First, while individuals are sensitive to how variants are distributed across individuals while they are learning, they do not condition their linguistic output on social variables. Indeed, there are no social variables in the model on which linguistic signals can be conditioned. The measure of bilingualism in this model comes from counting the discrete, symbolic, internal representations of individuals. The phenomenon that this top down model explores, then, is not the same kind of emergent population-level phenomenon that this thesis argues bilingualism must be.

Given their different assumptions, the models in this chapter reach different conclusions than other top-down models presented in the previous chapter. For instance, in model 2, bilingualism is inevitable for a wide range of settings. This stems partly from the fact that this model is not fully rational (see last chap-

ter), but also that this model was set up to focus on bilingualism. While this approach is not more ‘correct’ than others, it demonstrates that a researcher’s research bias can affect the conclusions that come out of models. The next chapter will demonstrate that a bottom-up approach can lead to further differences.

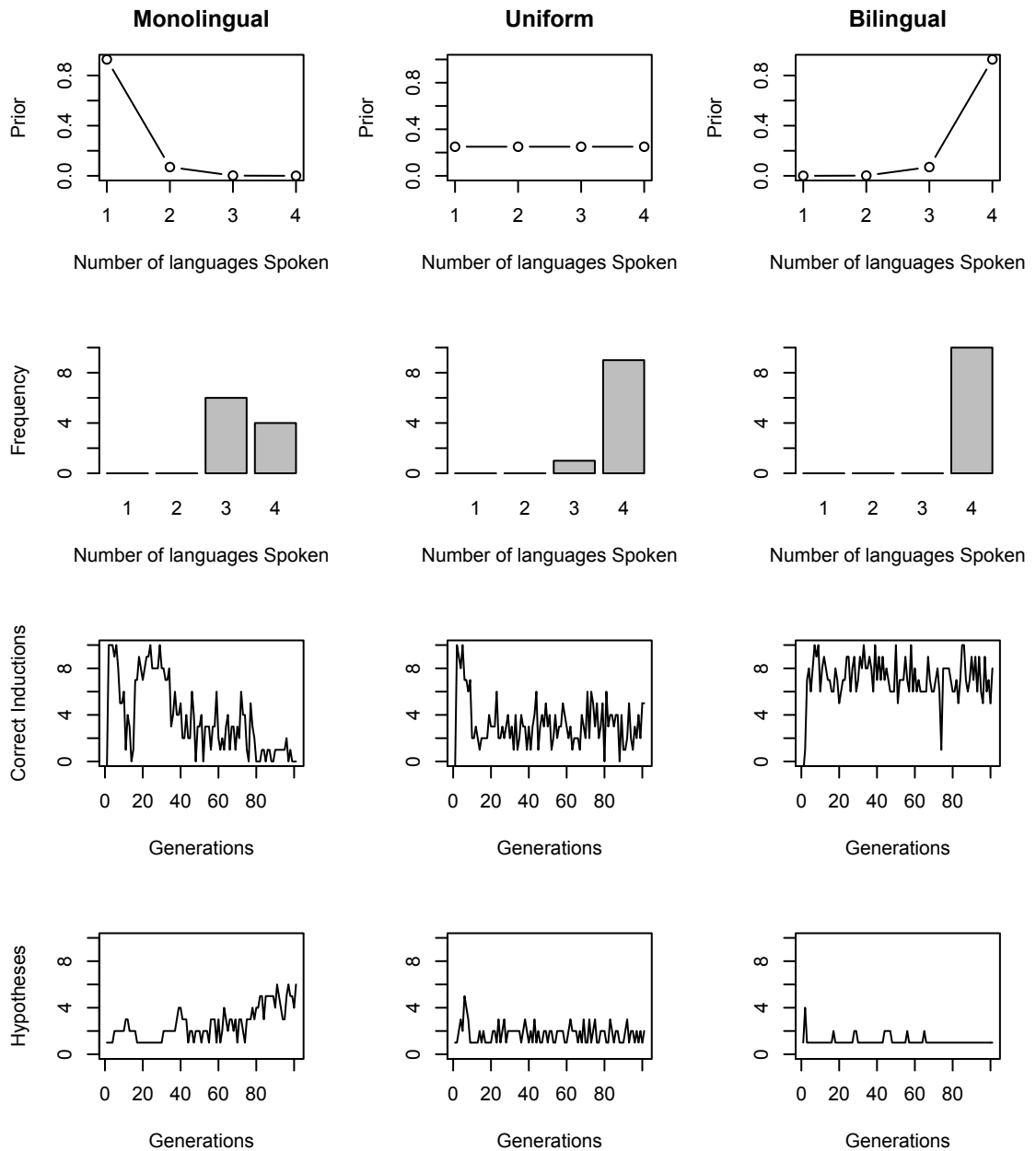


Figure 6.17: Results for a population of bilingual learners with different prior biases. Even with strong biases favouring monolingualism, bilingualism emerges. Rows, top to bottom: The distribution of the prior bias over hypotheses; Distribution of the number of languages spoken by agents in the last generation; Proportion of correct guesses by learners over generations; Number of hypotheses in the population. Columns, left to right: Monolingual bias, uniform bias, bilingual bias.

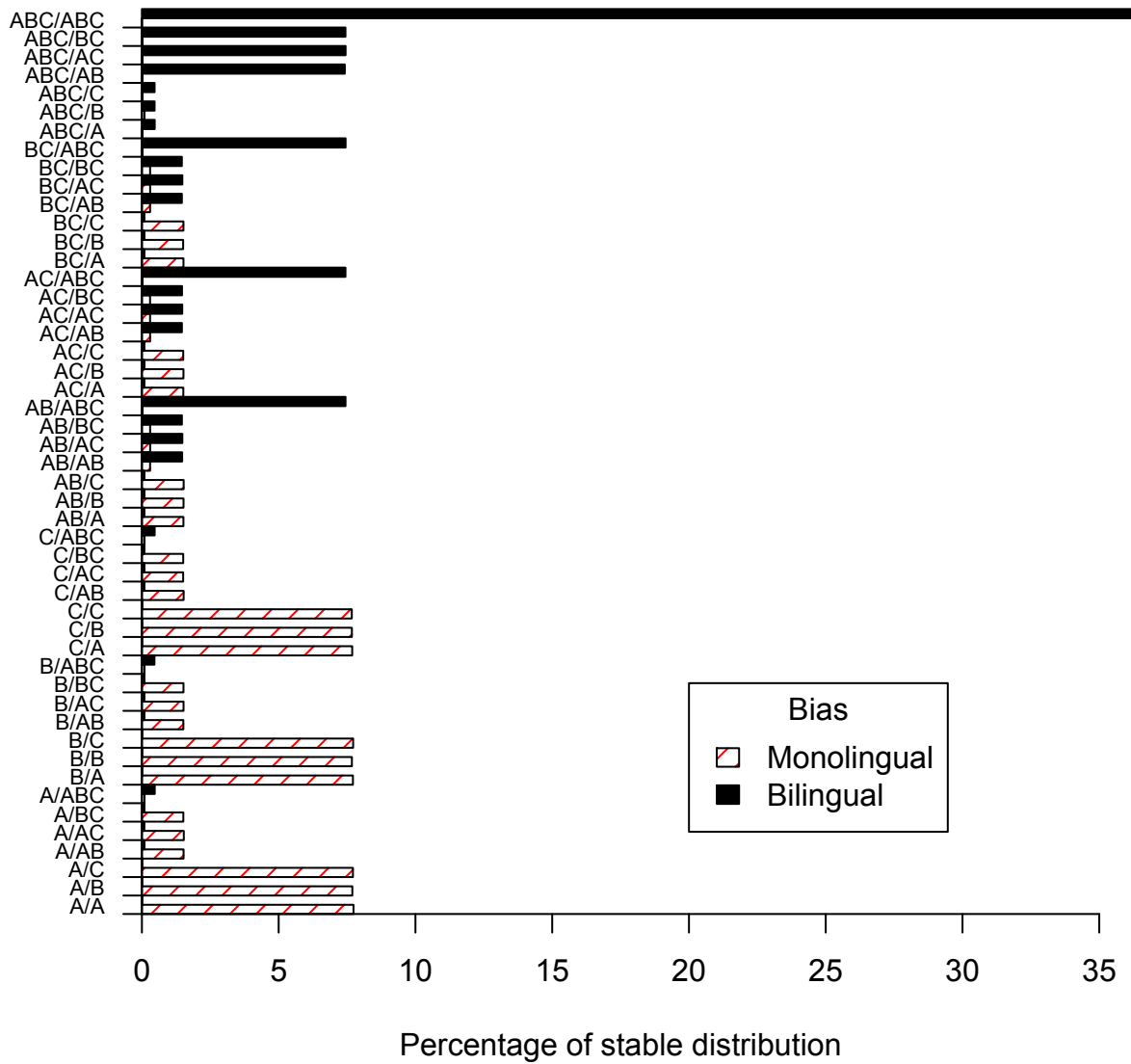


Figure 6.18: The stable distribution of contexts for a system with three languages, with two individuals in each generation and where each learner has two teachers. The hypotheses are shown on the vertical axis. The hypothesis of each teacher is divided by a forward slash. For example, ABC/B means teacher 1 knew languages A, B and C and teacher 2 knew only language B.

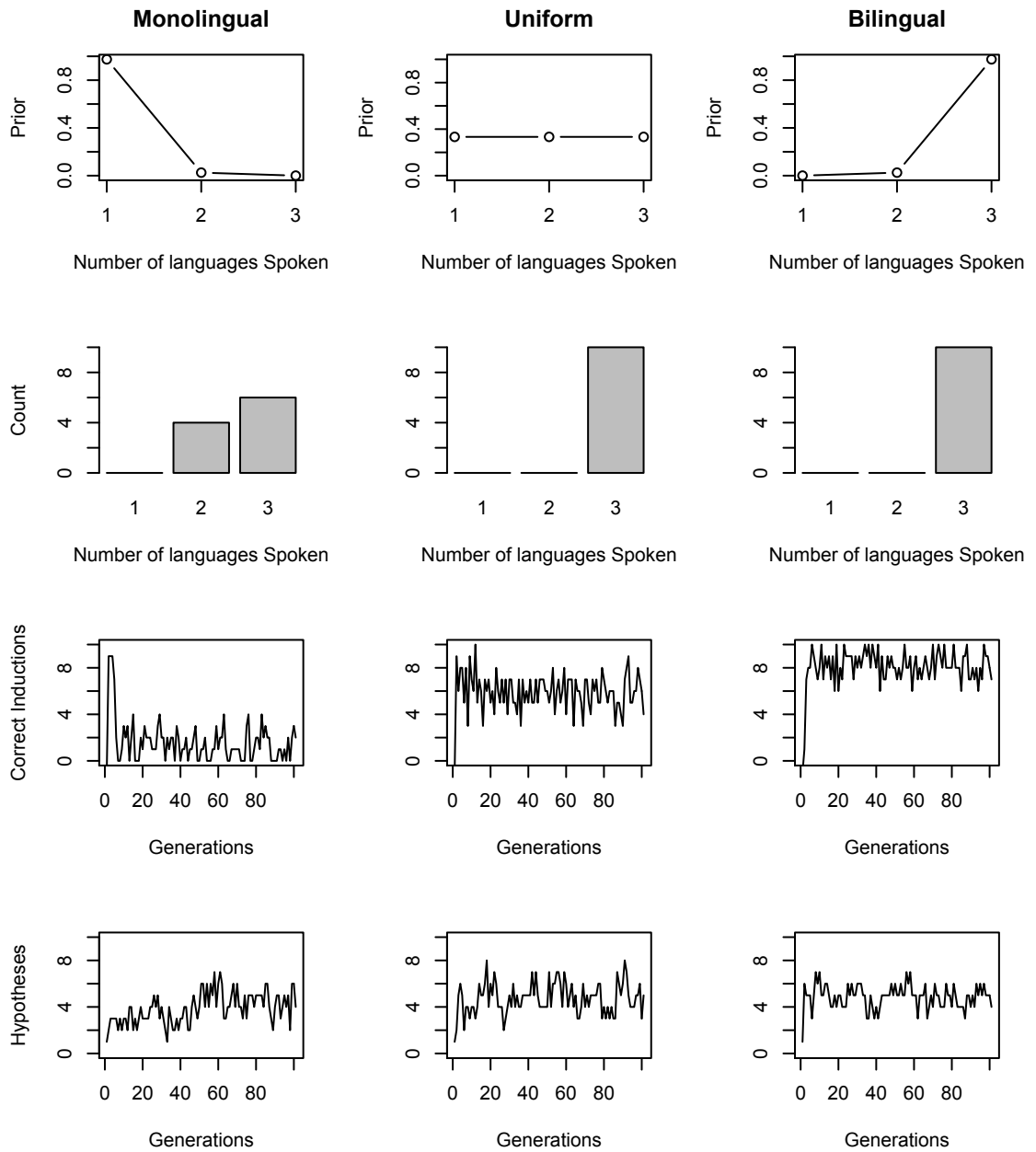


Figure 6.19: Results for a simulation with memory limits lower than the total number of possible languages. Rows, top to bottom: The distribution of the prior bias over hypotheses; Distribution of the number of languages spoken by agents in the last generation; Proportion of correct guesses by learners over generations; Number of hypotheses in the population. Columns, left to right: Monolingual bias, uniform bias, bilingual bias.

CHAPTER 7

A BOTTOM UP MODEL OF BILINGUALISM

“I realised that what I had said at sometime may have overemphasised rationality or some type of thinking and I don’t want to overemphasise rational thinking on the part of humans.”

John Nash (in Curtis, 2007)

7.1 Introduction

The last chapter presented a top-down, rational approach to the cultural evolution of multiple languages. The rational framework led to assumptions like discrete languages, fixed social structures and a specific learning mechanism for dealing with multiple languages. This chapter presents a bottom-up approach to the same question but with different assumptions: The linguistic signal is encoded as low-level and continuous; the linguistic signal is meaningful; social structures are dynamic and there is a general learning mechanism which has no biases over ‘languages’. The model is not free of biases, but they relate to general statistical properties of low-level features, unlike the top-down model which had built-in biases that were specifically designed to address the bilingual paradox. Bilingualism, as measured by the concrete definition developed in chapter 1, emerges in this model due to dynamic social structures. The results suggest that unconditioned variation is unstable, and that linguistic diversity tracks social change.

Taking a bottom-up approach means that there is no direct way of measuring the number of ‘languages’ that a speaker or community knows. Therefore, a bottom-up definition of bilingualism is required. By constructing this model, relevant measures of the variation in the linguistic signal will clarify what is meant by ‘bilingualism’. Essentially, this will be based on the difference between the amount you understand of another speaker’s total variation, and the amount you understand of another speaker’s variation when they are trying to communicate with you. When these are not equal, the situation may be interpreted as being bilingual. That is, bilingualism is a measure of how variation within speakers is conditioned on social variables. This is the measure of bilingualism described in chapter 1. The model will explore how variation changes as it is transmitted over generations in dynamic social structures.

7.2 Bottom-up approaches

Top down and bottom up approaches focus on different levels of explanation. David Marr's levels of explanation include the 'computational' characterisation of the problem, an 'algorithmic' description of the problem and an 'implementational' explanation which focuses on how the task is actually implemented by real brains (Marr, 1982). Top down approaches focus on the computational level of explanation, while bottom up approaches usually start with an algorithmic or implementation level of analysis.

A recent debate published in *Trends in Cognitive Sciences* discussed these two different approaches to cognitive science (McClelland et al., 2010; Griffiths et al., 2010). The central issue is which approach is the most productive for explaining phenomena in cognition. Proponents of the top down, 'structured probabilistic' approach argue that it is better suited to answering questions about how much information is needed to solve a problem, what representations are required and what the constraints on learning are (Griffiths et al., 2010). Griffiths et al. argue that using hierarchies of structure means that the model can be influenced by high-level information. For instance, a speaker may have a different attitude to a word if they are told its source language. Griffiths et al. argue that (bottom-up) connectionist models can't incorporate high-level information so easily, and can't 'change their minds' based on little data. They also argue that structured approaches can separate parts of a cognitive problem, for instance learning the structure and strength of a cause and effect (Griffiths and Tenenbaum, 2005). Connectionist models combine these two aspects, and it is often difficult to interpret at the computational level how a connectionist model is solving a problem (Eppler, 1993; Touretzky and Hinton, 1988; Intrator and Intrator, 2001). The advantage of the top-down Bayesian models presented in the last chapter is that the assumptions about the biases in the model are clear, and so it is relatively easy to explain the cause of a phenomenon in a model.

Proponents of the bottom-up approach make two counter-claims (McClelland et al., 2010). First, the 'top-down' approach is in danger of building in representations and structures into a theory that are not compatible with an implementational or algorithmic level of explanation. Instead of making claims about how cognition is structured, emergentists argue that structure should be allowed to emerge from the interaction between the data and a realistic processing mechanism. They note that the brain may not be solving problems optimally, as the structured approach assumes. Secondly, the emergentists argue that the structured approach cannot easily account for elements of development. Children exhibit learning curves and reversals in behaviour which can be captured by an emergent process, but not by a model that approaches the problem optimally. Some go as far to claim that Bayesian models are flexible enough to model almost any behaviour as optimal, and so are unfalsifiable (Bowers and Davis, 2012).

I agree with the Emergentists that assuming how the problem is structured is dangerous. Part of this thesis involves showing that structured approaches have made assumptions about the structure of the language learning problem that are not upheld when considering bilingualism¹. At the same time, emergentist models do make assumptions about the problem in the way they represent the input. For example, the input to neural nets might assume particular linguistic features are available to the learner (e.g. ‘wickelfeatures’, Rumelhart and McClelland, 1985 or conceptual categories, Rogers and McClelland, 2004)

The distinction between top-down and bottom-up may be a false dichotomy, however. Indeed, McClelland et al. see both approaches as being focussed on statistical learning and suggest that results from bottom-up and top-down models can feed into each other (McClelland et al., 2010, p.350). Also, the iterated learning model has not been approached exclusively with top-down tools: Smith et al. (2003) use neural network models of language learners in an iterated learning framework. Furthermore, the approaches may not be mutually exclusive. Friston and Kiebel (2009)’s dynamic expectation-maximisation model of learning uses hierarchies of structured models, but they work like neural nets for learning and Bayesian models for production. The current model also has aspects of top-down and bottom-up approaches, as discussed in the next section.

The important feature of a model, for the purposes of this thesis, is how it enables us to think about the bilingual paradox. In this case, models are tools for exploring the problem rather than definitive descriptions of a theory. One contrast that will appear from comparing the bottom-up approach and the top-down approach to bilingualism is that top-down frameworks, like Bayesian iterated learning, bias the researcher towards thinking about bilingualism as a property of individuals while bottom-up approaches, like the one described below, allow thinking about bilingualism as a property of populations.

7.2.1 A ‘bottom up’ approach to the evolution of bilingualism

The term ‘bottom-up’ usually refers to models that try to implement a realistic cognitive machinery. The learning mechanism in the current model (linear re-

¹I also disagree with the claim of Griffiths et al. that structured probabilistic models can determine how much information is needed to solve a problem. For instance, take visual navigation problems where an agent must navigate home using visual sensors. Early approaches assumed that you needed an internal map of the environment, an indication of where you were and your uncertainty, representations of headings etc. Research proceeded to approach this problem from a very high-level. However, Zeil, Hofmann & Chahl (2003) showed that you could solve the problem fairly effectively by taking the pixel-by-pixel difference from an image taken at the current location and an image taken at the target location. Importantly, this only worked in real, noisy environments, not artificially sparse labs. This may explain why ants and bees can navigate accurately despite having small brains - there’s enough complexity in the world for low-level systems to utilise.

gression) is not a plausible cognitive mechanism at the implementational level. Instead it is a computational level mechanism, which is usually associated with ‘top down’ approaches. However, the goal of this model is not to explore how humans acquire multiple languages, but instead to explore linguistic variation between individuals in a cultural transmission system. This model does take a bottom-up approach to the phenomenon that it is designed to explore: bilingualism. That is, its assumptions are designed to meet a concrete definition of bilingualism. While the top-down models in the last chapter had high-level representations of bilingualism (speaking multiple discrete languages) built into them, the current model only represents low-level linguistic variation. This variation becomes structured around social variables, so that ‘languages’ emerge from the ‘bottom up’.

In this sense, the Bayesian models presented in the last chapter can be seen to take a ‘bottom-up’ approach to the phenomenon of cultural transmission. The models are used as ‘null’ models which demonstrate the kinds of phenomena that can emerge solely due to cultural transmission through a bottleneck (Kirby, 2000; Smith et al., 2003; Kirby et al., 2007). The assumptions about the cultural transmission process in the current model are borrowed from these approaches. That is, the model is designed to demonstrate the minimal conditions under which bilingualism emerges. This chapter argues that a distinction between individuals and their linguistic systems and variation in social structures are the key features, while learning mechanisms that address the problem of learning multiple languages explicitly are not.

So, the current model is a mixture of techniques used in studies that are broadly categorised as ‘top-down’ and ‘bottom up’. However, the intended purpose of the model is to demonstrate that a top-down assumption about the problem - that learners must acquire discrete languages - biases researchers towards invalid conclusions about what is required to explore linguistic diversity in models of cultural evolution.

7.3 Model definition

This model is an iterated learning model (Smith et al., 2003) of the cultural transmission of linguistic signals. The model contains learners who are exposed to a series of communicative events. Each event involves a ‘teacher’ describing some semantic data with a linguistic signal. A ‘learner’ observes the event and must induce a model that predicts the linguistic signal given the semantic data. Perhaps a more accurate term for the ‘teacher’ is a ‘model’, since they provide data for the ‘learner’ to observe rather than actively engaging in ‘teaching’. However, this will become confusing below because each individual has a linguistic ‘model’ and ‘model’ also refers to the system in which the simulations take place, so I will refer to the individuals as ‘teachers’ and ‘learners’.

The learner’s learning mechanism is based on a stepwise linear regression of a set of continuous semantic variables that explain the variation in a continuous linguistic signal. A stepwise regression is a process of determining the best fitting model of the data by testing the fit of models with greater or fewer variables, according to an information criterion. Appendix A.1 has a brief introduction to stepwise linear regression.

A learner receives data as a sequence of measurements of various variables. The dependent variable is the linguistic signal. This is a real number that can be thought of as a representation of an utterance. The independent variables represent semantic features. These, too, are real numbers. These can be thought of as properties of the environment that the linguistic signal might be describing, or semantic properties of the sentence being described. Examples of the second instance may be gender, number or tense. Some real languages code for some of these at lower levels in the linguistic structure than others, for instance a morphologically marked past tense rather than a lexically marked past tense (Dahl and Velupillai, 2011). Some languages may not code for some properties at all, for example English does not encode different levels of politeness in its pronouns, but many languages do (Helmbrecht, 2011). The model is set up so that any semantic property may potentially become an important factor.

The environmental semantic properties are sampled from a specified distribution. Each semantic variable has a set of hidden parameters which describe the distribution from which values are sampled. This systematic variation is important so that the linguistic signal has some structure to emulate, as described in section 7.4.2.

Importantly, in addition to environmental semantic variables, the communicative event also includes the identity of the teacher who described the event. Depending on how the social structure is set up, this identity might be unique to each individual, or mark the identity of the community to which the teacher belongs (see section 7.3.1). This property is always available to the regression, but may not always be used.

The learning process is an iterated one: Learners receive a sub-sample of the data in the community, tagged with speaker identity. The probability of receiving a sample from a particular speaker depends on the structure of the community. After receiving the sub-sample of data, the learners perform a stepwise regression to settle on a model that fits the data. A new sample of the semantic variables is generated (new events occur) and the learner’s model is used to predict the linguistic signal based on these (new linguistic productions). This data becomes the input for the next generation, after being distorted by a small amount of noise. A noise parameter controls the level of noise that is applied. All individuals are replaced at each generation (there is no overlap of speakers between generations).

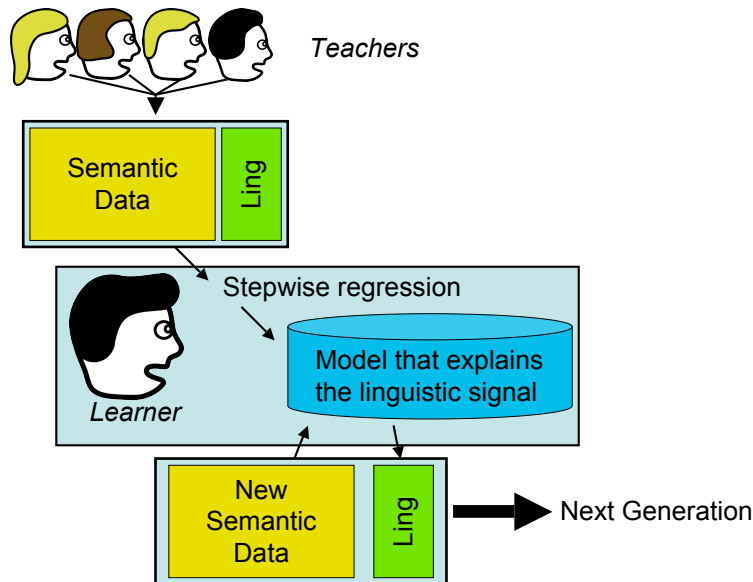


Figure 7.1: A diagram of how the bottom up model works. Teachers produce linguistic and semantic data for a learner to observe. The learner uses stepwise regression to build a linguistic model, which is then used to describe some new semantic variables for the next generation.

This process repeats for many generations. A diagram of the model is shown in figure 7.1.

7.3.1 Population parameters

Generations of individuals are separated by discrete timesteps $t_1, t_2 \dots t_n$. A population of N_t learners in the current generation observe data produced by N_{t-1} teachers in the previous generation. The probability of any given learner receiving data from any given teacher is defined by an interaction matrix $W(t)$. A learner i will receive data from teacher j with probability $W(t)_{i,j}$.

Since the population size and community membership can change over generations, a method is needed to specify this matrix succinctly. Therefore, the matrix is derived from a function with various parameters. The first parameter is the number of communities N_t at a particular generation. For each generation, a set $C(t)$ of N_t discrete labels represents which community each individual belongs to. This community feature is observable to the learners. In a generation with 2 individuals with 2 communities, $C(t) = \{A, B\}$. This means, for example that the first individual belongs to community A. Note that the community identity of different individuals may be perceived to be the same by learners.

The interaction matrix for individuals $W(t)$ is calculated from an interaction matrix between communities $I(t)$. The probability of learner i receiving data from teacher j is proportional to the interaction weight between the community

that i belongs to and the community that j belongs to:

$$W(t)_{i,j} = \frac{I(t)_{C(t)_i, C(t-1)_j}}{Sum_W} \quad (7.1)$$

Where Sum_W is the sum of all weights between individuals.

$$Sum_W = \sum_i^{P_t} \sum_j^{P_t} W(t)_{i,j} \quad (7.2)$$

The community structure can therefore reflect situations from simple ones such as ‘there are two communities’ to a weighted, directed graph between individuals.

The community interaction matrix $I(t)$ can be simplified to a vector of single numbers by assuming that the probability of receiving data from any community that a learner does not belong to is equal.

$$W(t)_{i,j} = \begin{cases} \frac{I(t)_i}{Sum_W} & \text{if } C(t)_i = C(t-1)_j \\ \frac{1-I(t)_i}{Sum_W} & \text{otherwise} \end{cases} \quad (7.3)$$

This assumption will be adequate for the examples in this chapter, and allows manipulation of the social structure through a single parameter for each community.

Here are some examples of social structures that can be set up. Given a situation where there are two teachers and two learners $N_{t-1} = N_t = 2$ and two communities at each generation $C(t) = C(t-1) = \{A, B\}$, different settings of I can then result in many social dynamics. Below I give some examples of matrices, with the learners (rows) labelled as L_1 and L_2 and the teachers (columns) labelled as T_1 and T_2 . For example, a society with two communities that are completely integrated and balanced (effectively a single community):

$$I(t) = \{0.5, 0.5\} \longrightarrow \begin{array}{cc} & T_1 & T_2 \\ L_1 & 0.5 & 0.5 \\ L_2 & 0.5 & 0.5 \end{array} \quad (7.4)$$

In the matrix above, for example, learner 1 (L_1) has an equal probability of receiving data from either teacher. Two communities that are completely isolated:

$$I(t) = \{1, 1\} \longrightarrow \begin{array}{cc} & T_1 & T_2 \\ L_1 & 1 & 0 \\ L_2 & 0 & 1 \end{array} \quad (7.5)$$

In the matrix above, learner 1 only receives data from teacher 1 and learner 2 only receives data from teacher 2.

The prestige of a community can also be manipulated. Below is the matrix

for a society with a majority and minority community, which has one community that receives input from both communities (the minority) and a community that only receives input from one community (the majority)²:

$$I(t) = \{0.5, 1\} \longrightarrow \begin{array}{ccc} & T_1 & T_2 \\ L_1 & 0.5 & 0.5 \\ L_2 & 0 & 1 \end{array} \quad (7.6)$$

Another type of transformation can involve distance, for instance assuming that the learners are spatially distributed over a single dimension, we can define the weights according to the distance between them, modulated by a factor d :

$$W_{i,j} = \frac{1}{|i - j|^d} \quad (7.7)$$

The equation above can be multiplied by equation 7.3 to combine spatial and social factors. More complicated social structures are possible, but these broad types will suffice for now. By manipulating C and I over generations, changes in social dynamic can be modelled such as migration, isolation, integration and social cohesion (majority vs. minority).

7.3.2 Measuring bilingualism

Since ‘languages’ are not encoded in the model, the amount of bilingualism must be calculated from the bottom up. There are two measures of learners’ linguistic similarity that are based on mutual intelligibility. The first is a measure of **comprehensive mutual intelligibility**: the proportion of utterances that one learner typically produces that another understands. For example, a monolingual speaker of English understands half of the languages spoken by a balanced bilingual speaker of English and Welsh. In the model, this is a measure of the proportion of the variance in one learner’s productions that is explained by another learner’s model. If we’re comparing individual A and B , this is implemented in the following way:

1. Generate semantic data in the same way as data was generated for the input.
2. Ensure that all speaker identities are sampled evenly in the data.
3. Get A to predict the linguistic signal for these data, based on A ’s linguistic model.

²It is noted that, in a closed population, imbalances in communicative interactions between communities is only possible if there is an imbalance in the size of communities. However, the imbalances used above can either be seen as modelling non-communicative interactions (such as input from television), or the idea that not all input becomes ‘intake’ that can be learned from (see Schmidt, 1990; Smith, 1993).

4. Get B to predict the linguistic signal for these data, based on B 's linguistic model.
5. Calculate the correlation between the two sets of linguistic signals.

See figure 7.2 for a diagram. If two learners have the same model, then they should be able to account for all of the variation in each other's productions (as the sample size approaches infinity). Individual A with a very different model from individual B will produce linguistic signals with a variation that is poorly explained by learner B's model.

We can also define a **functional mutual intelligibility** score which is the proportion of utterances that interlocutors understand when they design their utterances for each other. That is, a bilingual speaker of English and Welsh and a monolingual speaker of English could always make themselves understood by using English. In the model, this is calculated in a similar way to the comprehensive mutual intelligibility score, but the speaker identity is fixed to the identity of the speaker receiving the data:

1. Generate semantic data in the same way as data was generated for the input.
2. Set the speaker identity in the data to B 's identity.
3. Get A to predict the linguistic signal for these data, based on A 's linguistic model.
4. Set the speaker identity in the data to A 's identity.
5. Get B to predict the linguistic signal for these data, based on B 's linguistic model.
6. Calculate the correlation between the two sets of linguistic signals.

See figure 7.3 for a diagram. In this case, an individual with a linguistic model that used speaker identity as a conditioning factor would adjust its variation to better suit its receiver (i.e. in the Welsh-English example, by speaking only English).

This measure has a lot in common with Bell (1984)'s concept of audience design. Bell notes that variation within the speech of an individual can be due to linguistic factors, such as phonological effects, or due to the speaker designing their speech for their interlocutor. The functional intelligibility measure captures this affect of audience design. This will be discussed further in section 7.7.2.5.

The two intelligibility measures can be combined to get a measure of bilingualism

by subtracting the comprehensive intelligibility from the functional intelligibility.

$$\text{Bilingualism} = \text{FunctionalIntelligibility} - \text{ComprehensiveIntelligibility} \quad (7.8)$$

This can be calculated for the whole community by taking the mean bilingualism score for all pairs of speakers.

$$\frac{2}{n^2 - n} \sum_{i=0}^{n-1} \sum_{j=i+1}^n \text{Cor}(L(M_i, G(E, j)), L(M_j, G(E, i))) - \text{Cor}(L(M_i, E), L(M_j, E)) \quad (7.9)$$

Where $L(M,E)$ is a function that takes a speaker's linguistic model M and semantic variables E and produces a linguistic signal, and $G(E,x)$ is a function that takes semantic variables E and changes the speaker identity to x .

If the community bilingualism measure is near zero, this means that the comprehensive and functional intelligibility are similar, which means that communities are using the same medium, with possible optionality. A sociolinguistic analysis of this community might interpret this as a single language (monolingualism).

If the score is negative, the functional intelligibility is lower than the comprehensive intelligibility. For example, in the functional measure, speaker A would adapt their linguistic signal for speaker B and B would adapt their linguistic signal for speaker A. This yields a low functional similarity. However, their comprehensive similarity is high (their overall linguistic system is similar), so the bilingualism score is negative. This would be interpreted as bilingualism in the sense that each community has a medium they use within their community, but they can also use the medium of the other community to some extent. A negative bilingualism score means 'more' bilingualism in the lay sense. This is meant to represent the amount of linguistic variation that is conditioned on social variables, and so is analogous to an entropy-like measure where lower values indicate more order (the linguistic system is more conditioned on social factors = bilingualism) and higher values indicate more disorder (the linguistic system is less conditioned by social factors = monolingualism).

If the score is positive, the comprehensive intelligibility score is lower than the functional intelligibility score. For example, A adapts their linguistic signal for B, but B does not adapt their linguistic signal for A. This leads to a high functional intelligibility, but a low comprehensive intelligibility (A has a medium for use in their own community). This means that both communities share one medium, but one community has at least one other medium. This might be interpreted as a minority situation in which one community speaks a minority language as well as the majority language. As we will see below, it is useful to be able to distinguish between 'balanced' communities (negative scores) and 'minority' situations (positive scores).

Comprehensive intelligibility

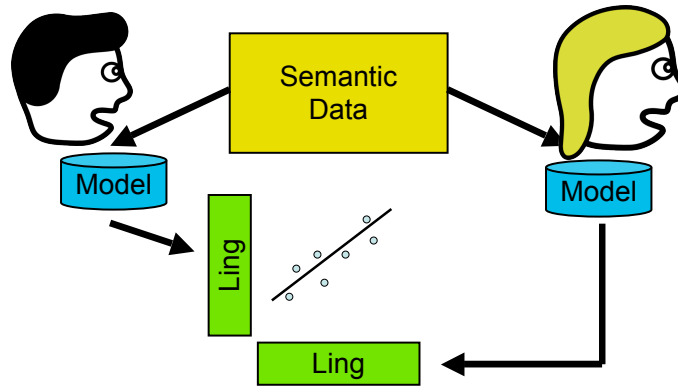


Figure 7.2: A diagram of how the comprehensive intelligibility measure is calculated. Two individuals are given the same semantic data and produce a linguistic signal with their linguistic models. The correlation between these signals is measured.

Functional intelligibility

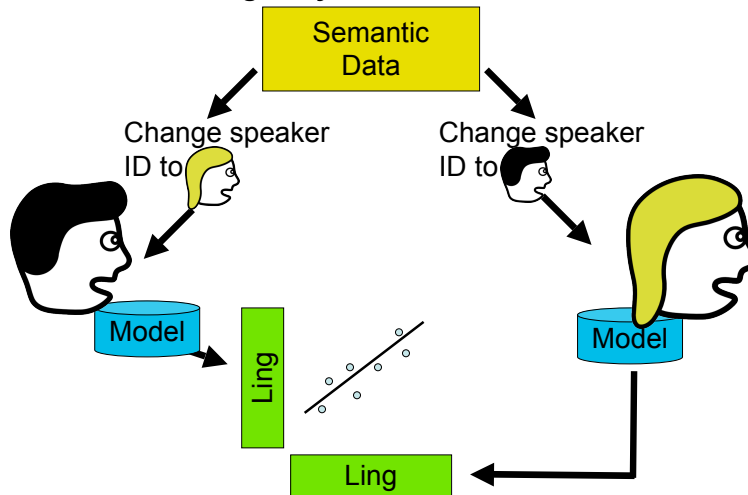


Figure 7.3: A diagram of how the functional intelligibility measure is calculated. Two individuals are given the same semantic data, but the speaker ID is changed to the other individual in the pair. They produce a linguistic signal with their linguistic models. The correlation between these signals is measured.

7.3.2.1 Applying the measure: Experimental example

One concern with the way the metric is calculated in the model is that it accesses the internal representations of individuals. To demonstrate that using the measure above is possible using non-internal measures, this section applies a bilingualism metric to some experimental results.

Gareth Roberts presents an experiment demonstrating the emergence of linguistic diversity in a laboratory experiment (Roberts, 2010b, Roberts, 2010a). This experiment was a game where individuals had to trade commodities in a series of rounds. At each round, individuals were paired up either with a partner in their group or outside of their group, though the speaker's true identity was hidden. Players were given random resources, but scored points based on how 'balanced' their resources were after trading. A commodity given to another individual was worth twice as much to the receiver as to the donor.

Players could only interact through an instant-messaging system. Prior to the game, individuals learned the artificial language that they were to use in these interactions. All participants were initially given the same starting language. The experimental conditions manipulated the frequency with which players interacted with their team-mate and whether the task was competitive or co-operative. In the co-operative condition, four players were considered as part of the same team and the task was to get as high a score as possible. In the competitive condition the four players were split into two groups and the task was to score more than the other team. In this condition, then, the main task was to identify whether your partner was a co-operator or a competitor. The only way to do this was through their utterances.

The results showed that, if players interacted frequently enough with their team-mates and were in competition with another group, then linguistic diversity emerged. Over the course of the game each team developed its own 'variety', and this was used as a marker of group identity. For example, in one game two forms of the word for 'you' arose. Players in one team tended to use 'lale' while players in the other team tended to use 'lele', meaning that players could tell group membership from this variation (it is advantageous to learn both your own group's variants and the other group's variants). Thus, linguistic variation arose due to the linguistic system evolving to encode the identity of the speakers.

We can measure the functional and comprehensive intelligibility of the participants in this study by looking at the lexicon in the following way. We'll assume that individuals understand words they use, and use all words they know at some point in the experiment. The comprehensive intelligibility of two individuals' utterances is the proportion of distinct words they use in common. The functional intelligibility is the proportion of distinct words that *A* uses when they think they are speaking to *B* that *B* also uses when speaking to *A* (players declared which

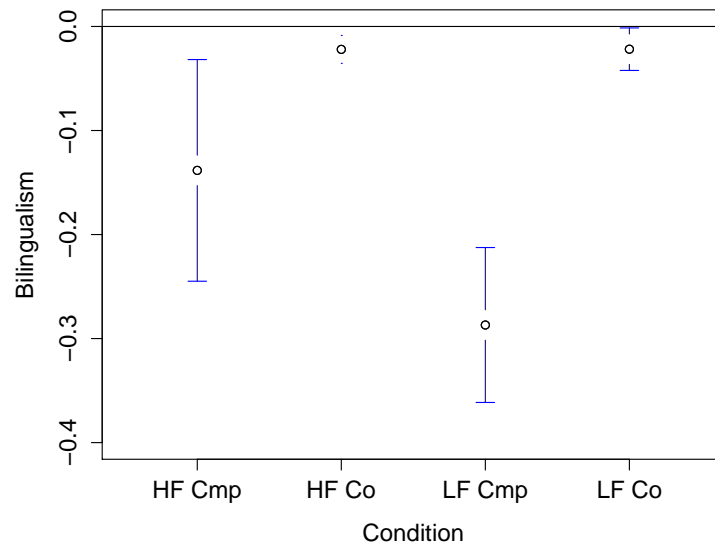


Figure 7.4: The bilingualism score for four conditions from Roberts (2010a). HF = high frequency (players interacted equally with their team-mates as other players), LF = low frequency (players interacted with other teams less than their own team), Cmp = Competitive condition, Co = Co-operative condition.

player they thought they were interacting with at the end of each round).

The mean functional and comprehensive intelligibility was calculated for each pair in the experiment. Figure 7.4 shows the results. An ANOVA revealed that there was a significant main effect of frequency of interaction with outsiders ($F(1,80) = 13.2, p = 0.002$), whether the individuals were in a co-operative or competitive condition ($F(1,80) = 52.3, p < 0.00001$) and a significant interaction between frequency and condition ($F(1,80) = 6.4, p = 0.02$). That is, the linguistic systems of individuals became more conditioned on the identity of interlocutors in the competitive condition than in the co-operative condition, and low frequency of interaction with outsiders accentuated this difference. This is in line with the results from Roberts (2010b).

Interpreting this for the purposes of my thesis, the competitive condition effectively introduced two community identities, while the co-operative condition only effectively had a single community identity and therefore fewer salient social variables to condition the variation on. The frequency of interaction with outsiders affected the social structure of the groups. Therefore, the results of this experiment show that community identity and social structure are important factors in the emergence of bilingualism.

In this thesis, I suggest that the level of bilingualism should be measured as

the amount of linguistic variation that is conditioned on social variables. There may be other ways of measuring this given the particular context of the study. For example, in the case of Roberts' experiment, we could use conditional entropy. That is, if a community is fully bilingual, then the linguistic system they use will be fully dependent on the person they are talking to. We can look at how the distribution of words that each speaker uses is conditioned on the person they are talking to (or think they're talking to). We divide the conditional entropy of the words they use given the identity of the individual they're talking to by the entropy of the sequence of words they use. This gives a measure between zero and 1. A high value indicates that there is little conditioning of the linguistic system on listener identity (with the extreme being full monolingualism) and a low value indicates that the words speakers choose to use are conditioned significantly on their interlocutor (with the extreme being full bilingualism).

Figure 7.5 shows the average conditional entropy for games by experimental condition. The results are broadly similar to the results for the functional/comprehensive bilingualism score. For instance, the order of conditions by mean values are the same. An ANOVA revealed, like for the bilingualism score, a significant main effect of condition (competitive vs. co-operative, $F(1,80) = 19.9$, $p < 0.0001$). However, unlike the bilingualism score, there was no main effect of frequency of interaction with outsiders ($F(1,80) = 0.5$, $p = 0.5$) nor an interaction between frequency and condition ($F(1,80) = 1.7$, $p = 0.2$). This suggests that, while the functional/comprehensive bilingualism measure is similar to entropy, it is not the same measure. The functional/comprehensive bilingualism measure may be more useful because it can tell the difference between bilingualism and a minority situation.

7.4 Interpreting the model

The data in the model are very abstract. Here is the most valid interpretation although, as I have shown for the Bayesian models, many interpretations are possible. Learners acquire a mapping between semantic events and linguistic signals. The semantic events are continuous and high-dimensional. The set of semantic variables is a sub-set of all possible semantic variables that the learner can perceive. In other words, the model assumes some sort of perceptual filter.

The linguistic signal is continuous, but one-dimensional. The difference between the production of the linguistic signal and its perception is modelled by introducing noise. While it might be easier to interpret this linguistic signal in phonetic terms (e.g. speech wave or vowel height), and difficult to imagine features such as syntax as continuous, I make no theoretical commitment. The linguistic variable could be limited to discrete numbers, or a binary variable without loss of generality. If the linguistic signal was represented by higher dimensions (e.g. by using a MANOVA or see the appendix for an example of a cluster learning mechanism

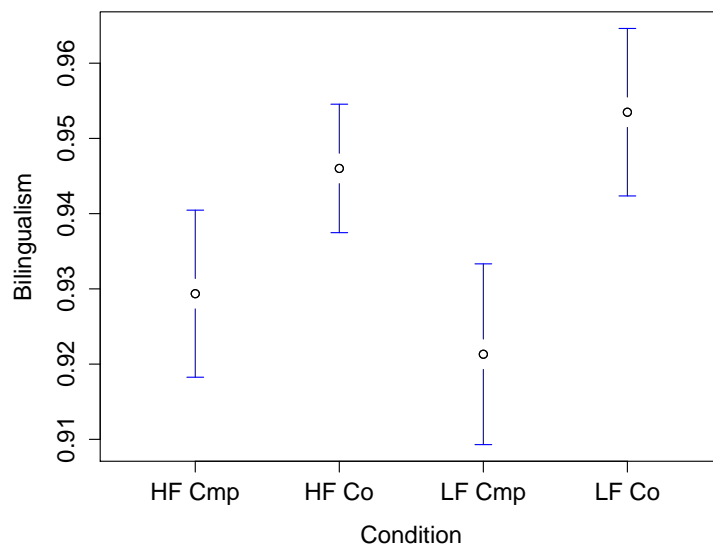


Figure 7.5: The average conditional entropy of speaker’s choice of words by the supposed identity of their interlocutor, from different experiment conditions from GARETH PhD THESIS.

that could do this), then any linguistic structure (or hierarchy of structures) could be represented. However, this model makes simplifying assumptions about these aspects in order to explore the phenomena of interest. The purpose of this model is not to investigate properties of the learning mechanisms of children acquiring language from real speech, but to illustrate a general model of cultural evolution in which the minimal requirements for a concrete measure of bilingualism can emerge.

Having said this, the difference between this model and many other models of iterated learning is that the mapping between signals and meanings is linear and continuous, not arbitrary and categorical. Due to the learner’s induction method being linear regression, certain mappings are not possible. For instance, synonymy is not possible (two different signals used to describe the same meaning) because a change in the signal necessarily means a change in the meaning. Homonymy is possible (the same signal used to describe two distinct meanings) if the slope of the regression is flat or if a learner’s model ignores changes to one semantic variable. These properties mean that the variation in the linguistic signal is correlated with the variation in the semantic signals. That is, the system is more indexical than symbolic.

However, this is a very strict interpretation of the model, and synonymy is not possible in the real world under this sense, because every signal suggests a slight difference in meaning (e.g. Clark, 1988; Wierzbicka, 1988). Mappings between

many signals and one setting of a particular semantic variable (a kind of synonymy) are possible: if we keep one semantic variable static, the linguistic signal might still change as a function of a change in another variable. More complex and realistic properties could be obtained by making the learning mechanism more complex. However, again, the object of this model is not to explore the mechanisms of learning, but of cultural transmission. While the interpretations of the data and learning mechanism may be ambiguous, the representation of the social structure and the measure of bilingualism are not. Learners are independent individuals who have their own linguistic model. They belong to a discreet community (whose size may be a single individual). They receive their data from other individuals in other communities based on a weighted graph (see section 7.3.1) which represents the social structure. The measure of bilingualism measures the amount of linguistic variation that is based on this social structure, which is an implementation of the concrete measure of bilingualism presented in section 7.3.2.

The model has two purposes: The first is as a proof of concept to show that an evolutionary model does not need to encode a discrete, static concept of a language in order to address questions about bilingualism. The model is theoretically valid by incorporating the properties above. In addition, the model should be empirically valid by being able to produce results that fit with findings in the field and be used predictively to extend other theories. Although obviously desirable, this is an ancillary goal for the purposes of this thesis. As will become clear, the model makes very general predictions, and ones that are not very surprising or new to the field of bilingualism. However, the model is supposed to illustrate a general theory of cultural transmission for the purposes of studying bilingualism.

7.4.1 Human function learning

While humans may not learn language using linear regression, it is at least obvious that humans are capable of learning linear functions. Kalish et al. (2007) run an experiment in human function learning. Here, participants were exposed to two co-varying variables whose relationship was defined by a function. The task was to learn the function and then predict the dependent variable given the independent one. Kalish et al. then take the independent variable values and predictions from the test phase of this participant and give it to the next participant as their input. Thus, an iterated learning chain is set up. Kalish et al. demonstrate that although trends in the responses of the first generation are unclear, there is a strong tendency for the chains to converge on a positive linear function, regardless of the initial function (see figure 7.6). This exposes a bias in individual learners for positive linear functions³. Kalish et al. also model this process by

³However, Ferdinand and Zuidema (2008a) run a similar experiment with two novel conditions. First, they tell the participants that some of the trials will contain random pairs (effectively telling the participants that the data will be noisy), and secondly using participants with mathematical backgrounds, so arguably with different priors. In both cases they find cases

using the Bayesian linear regression model discussed in section A.1.1. Figure 7.7 demonstrates that the current bottom up model can also simulate results similar to the human experiment results.

With regards to the iterated regression model discussed in this chapter, these results give some validity to using linear regression. Human learners are clearly capable of learning them and prefer them over some random and non-linear relationships. However, the question remains whether function learning is analogous to the kinds of problems faced by *language learners*. While there may be some analogues with iconic features of language such as sound symbolism (see Hinton et al., 2006), the general task of learning a language seems to be the formation of categories from data rather than function learning (e.g. realising that there are a certain set of discrete prototypical vowels from exposure to many instances of vowel sounds). Cluster analysis (see section A.2 on page 195) or topographic learning (which may also be more neurologically plausible, Ellison, 2012, see section A.3.1 on page 193) may offer a more parsimonious approach with regards to the mechanism of learning, but this is left for future exploration.

7.4.2 The role of the environment

In this bottom-up model, the environment is a crucial factor. The distribution of non-linguistic variables (the structure of the semantic space) will shape the distinctions that emerge in the language. Furthermore, variation or perturbation of the environment can keep the variance in the linguistic signal from reducing to nothing.

Perfors and Navarro (2011) show that the structure of the environment is indeed important in an iterated learning paradigm with human participants. Participants had to learn words for stimuli in a continuous meaning space defined by size and colour (shades of grey). The labels for each meaning were elicited from the participants and then these labels were passed on as the input to another participant. This was repeated to create an iterated chain of learning, much like (Kirby et al., 2008). The chains were split into two conditions with slightly different meaning spaces that favoured a division by size or colour respectively. The meaning space that favoured size had a steep gradient in the middle of the meaning space, leaving a salient categorical boundary. The meaning space that favoured colour had the same manipulation in the other dimension. Over generations, labels for different areas of the meaning space merged into one another, but respected the boundaries defined by their particular meaning space bias. That is, lexicons in chains with a size-bias meaning space tended to develop a big-small distinction, while lexicons in chains with a colour-bias meaning space tended to develop dark-light distinctions (see figure 7.8). They argue that assuming that a learner's probability distribution over languages is independent of the structure of the world may be unrealistic.

where nonlinear functions emerge.

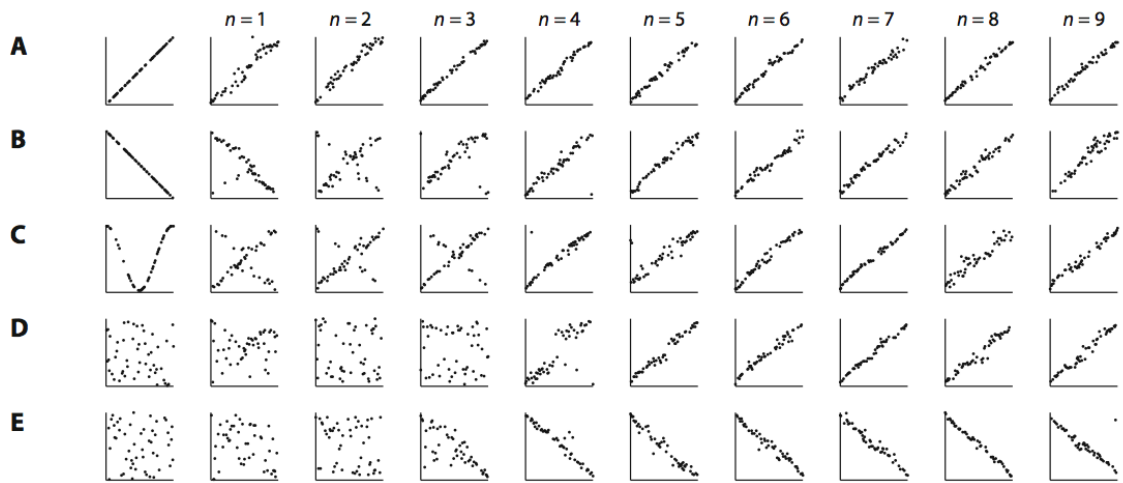


Figure 7.6: Results of the iterated function learning experiment from Kalish et al. (2007). Each graph shows the output of a particular participant as the relationship between the dependent and independent variables. Each row is an individual chain with each column representing subsequent generations. The first column is the initial input given by the experimenters. These follow a positive linear function (A), a negative linear function (B) a U-shaped function (C) and a random one-to-one mapping (D). Series E shows a case where a negative linear bias emerged.

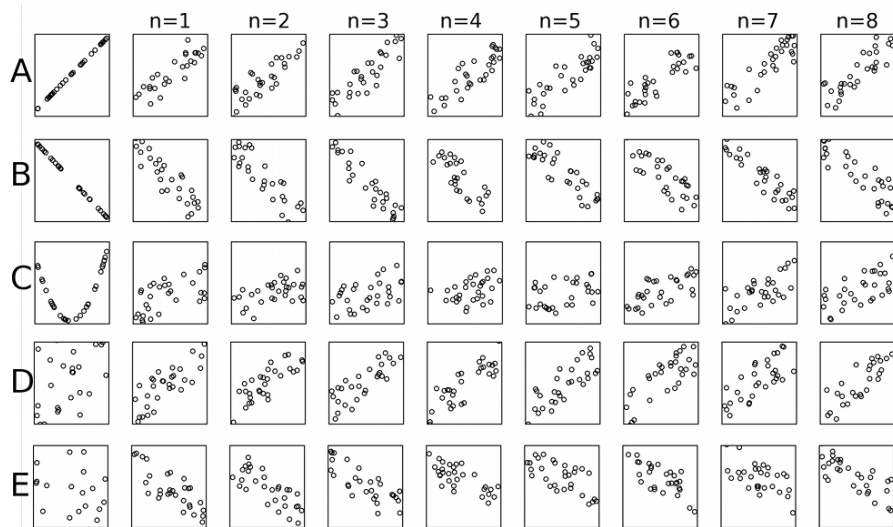


Figure 7.7: Results from the bottom up model simulating the results from Kalish et al. (2007), see figure 7.6. 1 speaker, 1 semantic variable with a uniform distribution, 30 data points, error = 0.5, information criterion $k=2$.

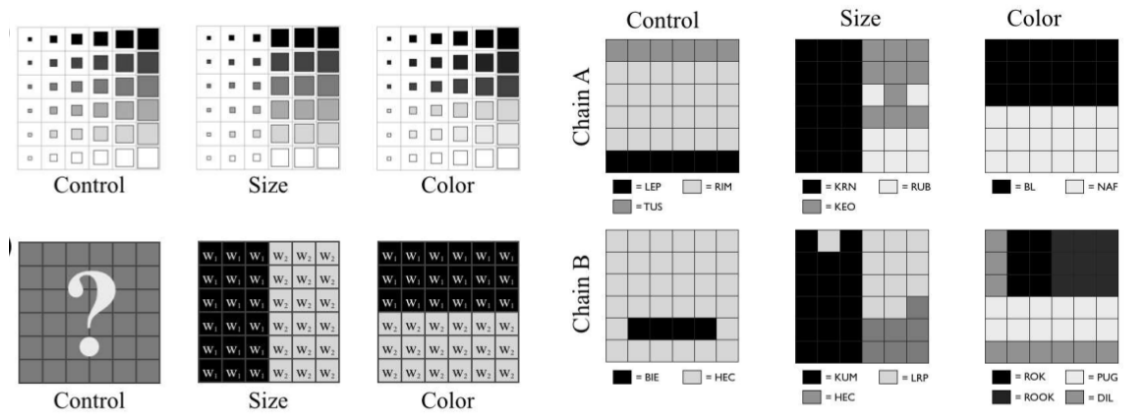


Figure 7.8: Stimuli (left) and results (right) from Perfors and Navarro (2011). Left, top: the different meaning spaces. Left, bottom: Predictions for the final outcome of the lexicon after iteration. Right: The lexicon at the 10th generation for six chains run in three conditions. In the two experimental conditions the lexicon has changed to reflect the structure of the meaning space.

The environment variables in the current bottom up model are very flexible. The only assumptions about the variables are that there is no multicollinearity (there is enough data), the observations are independent and that the distribution is homoscedastic (the variance is constant for a particular variable)⁴. The properties of the environment can change over generations, allowing the modelling of a changing environment. New variables can also be added between generations to model new innovations, or old ones removed from the analysis altogether to model a more dramatic change in the environment such as experienced during migration or conflict. This differs from a standard Bayesian approach which assumes that all learners have the same prior probability distribution over hypotheses (e.g. Griffiths and Kalish, 2007), or at least all learners have the same hypothesis space. This is difficult to reconcile with a realistic view of long-term language change where new conditioning factors are constantly emerging. While a Bayesian approach could account for this, it would have to change its definition of the learner’s learning process. Changing the environment in the current model is valid without changing the learner’s learning mechanism.

7.5 Biases in the model

The Bayesian learners in the top-down models have exactly specified prior biases. In the bottom-up model, there are no biases that are explicitly built in to the learner with parameters in the same way. However, there are restrictions on the space of models, attractors in the space of linear models in an iterated linear regression process and there is a built-in restriction against learners adopting models containing only the intercept. These are described below.

⁴The cluster analysis model also assumes gaussian distributions.

7.5.1 The restrictions of linearity

The use of a linear regression as the learning mechanism limits the mappings between the linguistic signal and the semantic variables that an individual can learn. This is effectively a prior bias over the possible mappings. While there are statistical mechanisms that can learn non-linear mappings (see the section on recursive filters A.3.5 on page 199), these are not considered here for simplicity. However, as mentioned in section 7.4.1 above, there is evidence that humans have a bias towards linear relationships (Kalish et al., 2007). It is possible that non-linear learning mechanisms could provide support for more bilingualism by being able to map more than one signal onto the same meaning. However, the purpose of the model is to explore the dynamics of bilingualism in a model of cultural transmission, not to explore cognitive learning mechanisms.

7.5.2 Attractors in the model space

The slope of a semantic variable in learners' models can change from generation to generation. In a situation with one individual at each generation, a learner (slope at time t) will adopt its teacher's slope (slope at time $t - 1$) exactly if there is no noise. However, in the presence of noise, the slope will change. The amount and direction of this change may be affected by other factors apart from the slope at the previous generation. This effect represents a *bias* in the model. For example, if there was no bias, then the slope difference between generations should be the same regardless of the value of the teacher's slope. Instead, the model exhibits a negative relationship between the teacher's slope (slope at $t - 1$) and the difference between the teacher's slope and the learner's slope (slope difference between generations). Recall that slopes can have a coefficient greater than one, but tend to be limited in the results below to within -1 to 1 due to the range of the semantic variables.

When the slope of the teacher is small, the slope change tends to be small and random (equally distributed above and below zero). However, as the slope of the teacher increases, the slope change decreases. That is, if the teacher has a slope of 0.5, the learner change between slopes is more likely to be negative than positive (the learner has a slope smaller than the teacher). This means the model has a bias against steep slopes. This is partly to do with the stepwise regression parameters and the distribution of semantic variables. Stronger biases (against big slopes) result from lower information criterion c value, higher noise levels and the degree to which the semantic variables are related (see appendix C).

The number of variables in a learner's model is affected by the information criteria c value, and tends to increase over generations (see appendix C). The dynamics of the model can be sensitive to initial conditions (the model of the first generation, see appendix C.6.1.4). If the first model includes some semantic variables, there will be no drastic change in how the balance in the use of the semantic variables

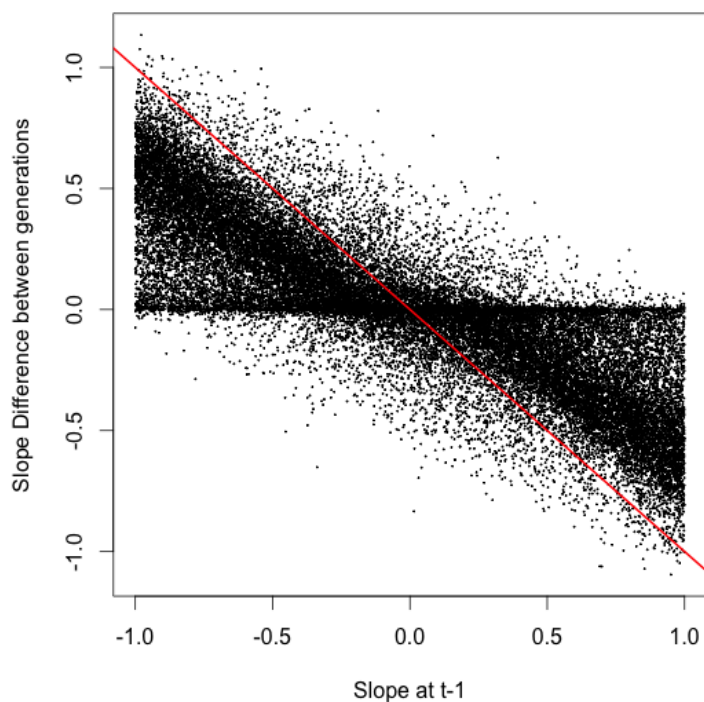


Figure 7.9: The relationship between the slope of the teacher (slope at $t-1$) and the slope change between generations. The red line has an intercept of 1 and a slope of -1 on this graph, indicating a perfect (negative) 1:1 relationship between the variables.

over subsequent generations. However, if the first model has no semantic variables, then the change in the balance between the two semantic variables will be considerable over generations.

Given differently distributed semantic variables, the linguistic models tend to select variables with smaller variances (see Appendix C.7). The model dynamics (regarding slope) are not affected qualitatively by the type of distribution of the semantic variables (e.g. normal, bimodal, uniform, see Appendix C.5). For this reason, semantic variables will have normal distributions in the main model.

Allowing an intercept decreases the strength of the bias for small slopes (see appendix C). The intercept allows a greater range of slope fits and so the models are more tolerant to larger slopes. For this reason, and to better fit existing linear modelling theory, the main model will include an intercept term.

Greater noise levels lead to fewer variables in models on average, steeper slopes and more balance between variable coefficients. However, these effects are small. Greater sample sizes lead to fewer variables in models on average and smaller intercepts, though these effects are small. However, smaller sample sizes do produce

shallower slopes on average.

7.5.3 Convergence to the intercept

There is a possibility that the best model that describes a learner's sample has only an intercept and no predictor variables. If the learner adopts such a model, then they will produce no variation in their output. Subsequent generations will therefore also be unlikely to use semantic variables, and the system will not evolve (although enough noise in the transmission could lead to a semantic variable being re-introduced, this is unlikely)⁵ In order to avoid this, learners only consider models with at least one variable⁶. This introduces another bias over the hypothesis space, although a very general one.

7.6 Example

This section walks through the steps of modelling a particular situation. For this example, we'll model two communities with two members each. There'll be one semantic variable.

7.6.1 Population Structure

We'll run the model for two generations (timesteps t_0 and t_1) with no changes to the population structure (two members in each community in both generations). Therefore $N_0 = 4$ and $N_1 = 4$ and the community structure is $C(0) = C(1) = \{A, B\}$. This model will be set up so that learners are three times more likely to receive data from their own community than from the other community. So, $I(0) = I(1) = 0.75, 0.75, 0.75, 0.75$. This yields the interaction matrix from

⁵This effect might parallel one encountered in the iterated learning paradigm (Kirby et al., 2008) where the variation in linguistic signals that are passed through a bottleneck tends to decrease without a counteracting pressure (e.g. for expressivity).

⁶There are two other possibilities. The first option is that the learner assumes that there must be *some* explanation for the linguistic signals they are hearing, so they lower their criteria for including variables in their model (i.e. presented with low variation, they choose a model with a single variable or interaction term that explains the greatest amount of variation in their input). The second option is that the learner assumes that if the best model contains only an intercept term, then there is no systematic behaviour, so they behave randomly. This could be modelled by producing data that was normally or uniformly distributed around the intercept with the same standard deviation as the linguistic input. This would prevent the variation from dropping rapidly and could create correlations at random that the next generation could pick up on. The more parsimonious option is the first, since it does not introduce any new dynamics into the system, apart from an adaptive reduction of the model selection criteria. It also fits better with a pressure for expressivity. However, this is not guaranteed to stop the convergence over time. Furthermore, if the information criterion is not necessarily constant, this limits the ability to analyse the affect of information criteria on the model dynamics.

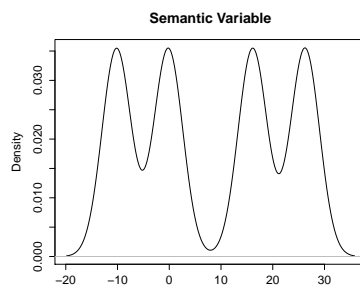


Figure 7.10: An example of a randomly generated multimodal distribution for a semantic variable, used in the example in section 7.6.

teachers T to learners L for t_1 as:

$$I(t) = \{0.75, 0.75, 0.75, 0.75\} \longrightarrow \begin{array}{cc} & \begin{array}{cccc} T_1 & T_2 & T_3 & T_4 \end{array} \\ \begin{array}{l} L_1 \\ L_2 \\ L_3 \\ L_4 \end{array} & \begin{array}{cccc} 0.75 & 0.75 & 0.25 & 0.25 \\ 0.75 & 0.75 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.75 & 0.75 \\ 0.25 & 0.25 & 0.75 & 0.75 \end{array} \end{array} \quad (7.10)$$

7.6.2 Parameters

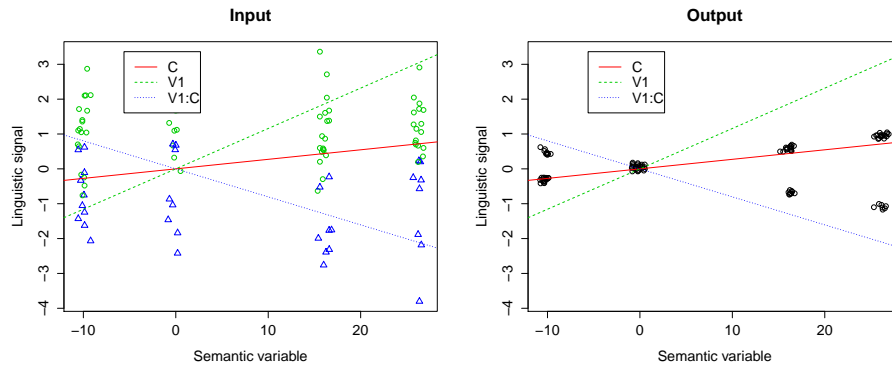
This example will use one semantic variable with a multimodal distribution. A set of randomly chosen hidden parameters specify the mean and standard deviation of 4 normal distributions. These combine to give a multimodal distribution (for an example, see figure 7.10). Each learner will observe and produce 100 utterances. The Akaike information criterion (AIC) will be used for the stepwise model fitting. The noise parameter will be set to 0.1. The semantic variable will be re-sampled between generations, and noise will be added to the linguistic signal between generations. The linguistic signal that the first generation receive is a uniform distribution over the range of the semantic variables. The community ID variable is distributed equally across the whole data between the two communities.

7.6.3 Procedure

First, the semantic variable is initialised (see figure 7.10). The first generation receive their data, which is a sub-sample of the total data available, sampled according to the interaction matrix. Individual L_1 's data is shown in figure 7.11. L_1 then fits a model to this data, which is also represented in figure 7.11. L_1 's model includes the semantic variable, the community variable and the interaction between the two.

After each learner has sampled data and fit a model, some measurements can be made. The most important is the measure of bilingualism. Equation 7.9 is applied by measuring the functional and comprehensive intelligibility between each pair of individuals and taking the average. In the first generation, this turns out

Generation 1



Generation 2

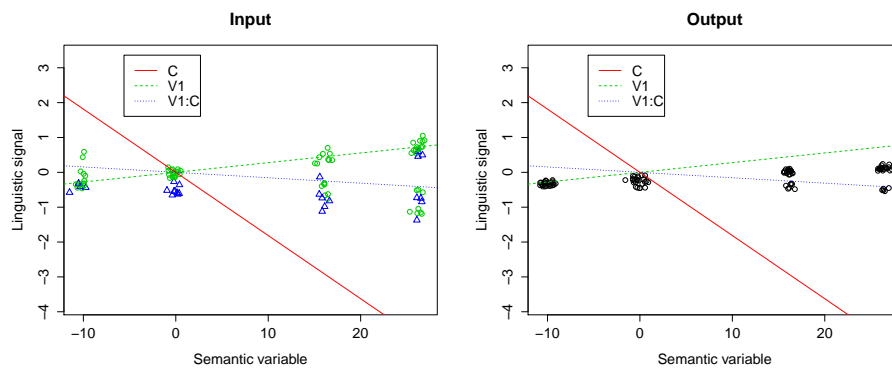


Figure 7.11: The input and output of two individuals in subsequent generations. The top row represents input (left) and output (right) for an individual in the first generation. The x axis of each graph represents the semantic variable and the y-axis represents the linguistic signal. Green circles indicate an utterances from the individual's own community and blue triangles represent utterances from the . The regression lines for each term in the learner's model are represented. C=community ID variable, V1 = semantic variable, V1:C = the interaction between the community ID and the semantic variable. The individual's output is calculated based on this regression model. The bottom row represents the input and output of an individual in the second generation.

to be around -1.3 (and climbs to about -0.6 in the second generation).

After this, the generation turns over. The learners become teachers, and each teacher produces 100 utterances, given new semantic data to describe. Individual T_1 's productions are shown in figure 7.11. These productions go on to be re-sampled by the next generation of learners.

7.7 Results: The cultural transmission of bilingualism in populations

The dynamics of the model are explored in appendix C. Here I present the basic findings of the model for different kinds of population structure with multiple speakers and multiple predictors. To summarise them briefly: unconditioned variation is unstable and bilingualism tracks social change.

First, the model and bilingualism scores behave as one would expect. When the two communities are completely integrated (integration parameter $I = 0.5$), then they quickly converge to using the same linguistic signals. The bilingualism score quickly converges to zero (figure 7.12). When the two communities are partially isolated (integration parameter $I = 0.8$), their varieties will take longer to converge. Negative bilingualism persists for a number of generations (figure 7.13). The results are slightly different in a minority situation where learners from one community receive input equally from both communities (the minority, integration parameter $I = 0.5$), but the other community mainly receives input from speakers from its own community (the majority, integration parameter $I = 0.9999$). In this case positive bilingualism emerges and is maintained for many generations (figure 7.14).

Figure 7.15 shows the relative importance of speaker rank in the three social situation scenarios⁷. In the integrated and isolated scenarios each community ranks speaker identity with the same importance. The speaker identity is initially somewhat important, but then becomes less important over time. However, the rank of the speaker identity in the integrated scenario changes quicker than in the partially isolated scenario. In the minority scenario, speaker identity remains relatively important for both communities, but is more important for the minority community.

These results are for communities with static social structures. We can manipulate the social structure to demonstrate that linguistic diversity also tracks the *change* in social structures. Figure 7.16 shows the results of simulations with dynamic social structures. The communities go through a cycle of being integrated,

⁷Calculated as the rank of the speaker identity variable by coefficient magnitude, ignoring interactions, divided by the number of parameters in the model.

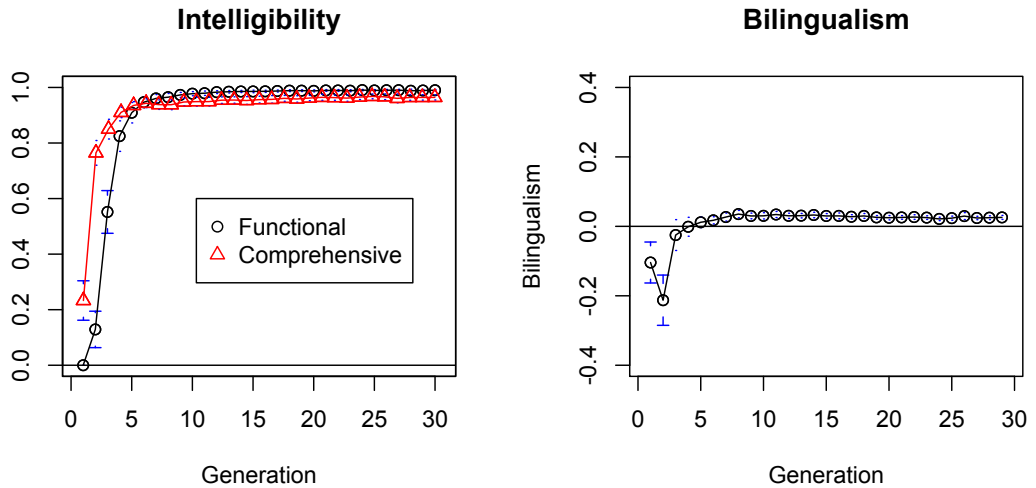


Figure 7.12: **Integrated.** Results for 100 simulations with 2 communities of 2 individuals at each generation where learners from one community receive input equally from both communities (the minority), but the other community only receives input from speakers from its own community (the majority). The graph on the left shows the functional (black) and comprehensive (red) mutual intelligibility scores. The graph on the right shows the community bilingualism score (functional intelligibility minus comprehensive intelligibility).

isolated, integrated and isolated again, with a few transition generations between each phase where the integration parameter is interpolated gradually. As shown above, if two communities are integrated, they will come to speak effectively a single medium (bilingualism score around zero, see figure 7.16). However, as the communities become more isolated, the bilingualism score increases. This is also in line with the results above. However, as the communities increase their interactions after this, the linguistic system changes again so that everyone speaks a single medium. Then we can split them apart and two varieties will emerge again with some amount of bilingualism. That is, the distribution of linguistic variation tracks the changes in social structure.

7.7.1 Factors that influence bilingualism

As is evident from the analysis above, the integration parameter affects the bilingualism score to a large extent. However, some further questions remain. For instance, why does negative bilingualism arise during the first contact phase, but then positive bilingualism emerge for the other contact phases?

Negative bilingualism is inherently unstable in this model. As soon as individuals start mutually accommodating the linguistic signal of other communities, this neutralises the distinction over speaker ID. This affect could be exaggerated because the model assumes that generations are discrete. However, it does not mean that bilingualism in the lay sense is unstable. Firstly, the bilingualism

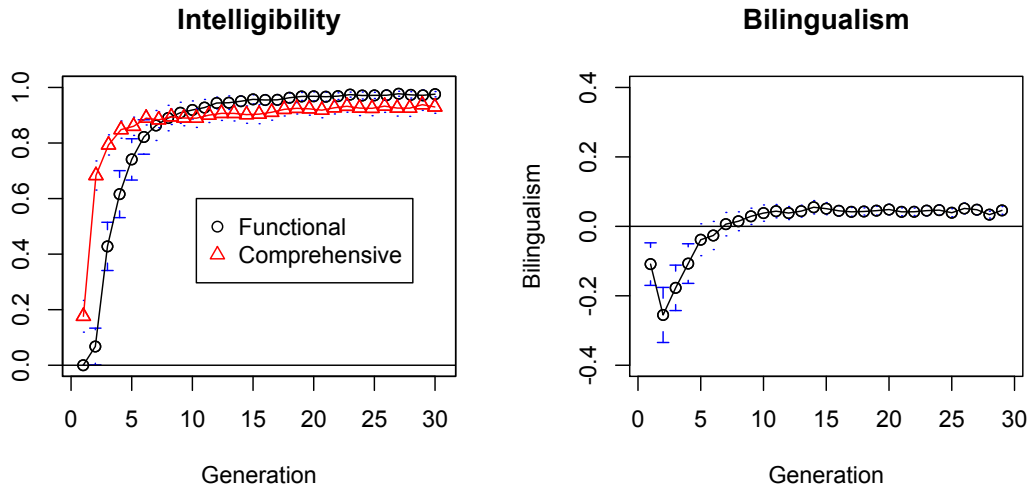


Figure 7.13: **Isolated.** Results for 100 simulations with 2 communities of 2 individuals at each generation where the integration parameter is set to 0.8 (learners are more likely to get input from speakers in their own community than the other) with three semantic variables and one hundred data points of input for each learner. The graph on the left shows the functional (black) and comprehensive (red) mutual intelligibility scores. The graph on the right shows the community bilingualism score (functional intelligibility minus comprehensive intelligibility).

score is not necessarily an index of an intuitive idea of bilingualism. A community like those in Catalonia might actually score zero on this bilingualism scale, because many people speak both languages. Secondly, in the real world, linguistic variation might be dictated by social factors not modelled here, such as location, formality or stage of the conversation (e.g. Labov, 1963; Meyerhoff, 2008). Finally, this model includes no pressures to maintain a linguistic identity such as prestige, politics or resistance to freeriders (Gareth Roberts, 2010b). Rather, it demonstrates the kinds of bilingualism that can emerge just from the process of cultural transmission - a kind of baseline behaviour on top of which more complex factors are applied.

The main determining factors of the bilingualism score are the importance of speaker identity as a conditioning factor and how differently each individual ranks speaker identity (see appendix section C.9.1). Negative bilingualism is much more likely to emerge if speaker identity is the most important conditioning factor, while positive bilingualism scores can emerge if speaker identity is less important. Negative bilingualism is also more likely if individuals rank speaker identity in their models similarly. Section C.9.2 discusses some more complex interactions. For example, negative bilingualism scores (as opposed to positive bilingualism scores) tend to emerge when: the speaker ID rank is low in the previous generation (more important), except when the integration parameter is increasing (communities diverging), when it can be higher; when the community with the largest model also considers speaker ID to be less important; when the

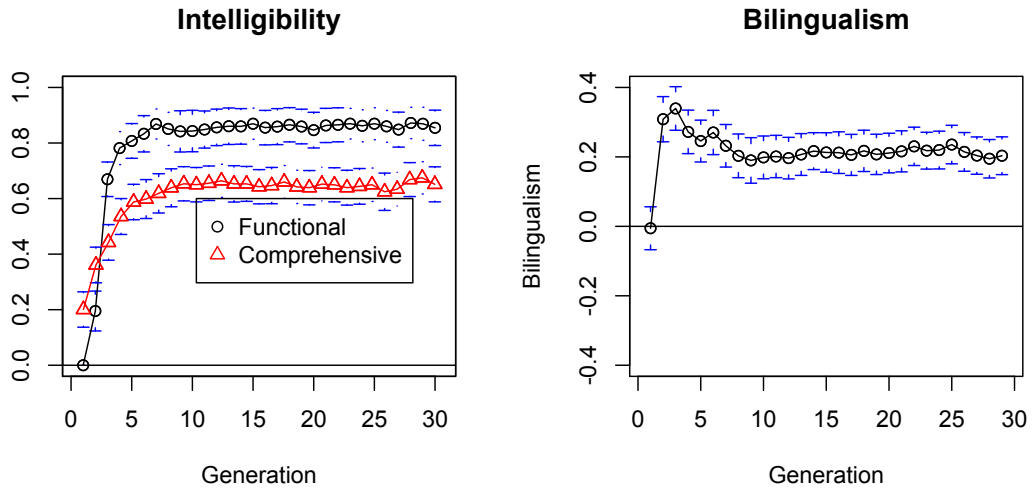


Figure 7.14: **Minority.** Results for 100 simulations with 2 communities of 2 individuals at each generation where learners from one community receive input equally from both communities (the minority), but the other community only receives input from speakers from its own community (the majority). The graph on the left shows the functional (black) and comprehensive (red) mutual intelligibility scores. The graph on the right shows the community bilingualism score (functional intelligibility minus comprehensive intelligibility).

mean and standard deviations of the speaker id rank are correlated; and when there is a stronger correlation between the difference in linguistic signal means and model fit ratio between communities.

In figure 7.16, after the first contact situation, only positive bilingualism tends to emerge. This is partly due to the linguistic signal of two communities adapting to the same semantic distributions, and so becoming more alike. Negative bilingualism requires that there are large differences in the linguistic signals of each community so that speaker identity conditions a large amount of variation. In a population with a positive bilingualism score there is an imbalance in the extent to which different communities adapt to each other’s linguistic signal. It is possible to identify a ‘superstrate’ community as the one whose linguistic signal changes least between the generations of contact. The comprehensive intelligibility score was used to measure the linguistic difference between generations (see section 7.3.2 above and appendix C.9 on page 240). A linear regression was run to determine what factors influenced this measure.

The difference in the linguistic signal means between generations is the main determiner of which community’s language will change the most. If community A’s mean is higher than community B’s mean in the previous generation, then community A’s language will change more than community B. This affect arises due to the bias in the model for small intercepts (see section 7.5.3).

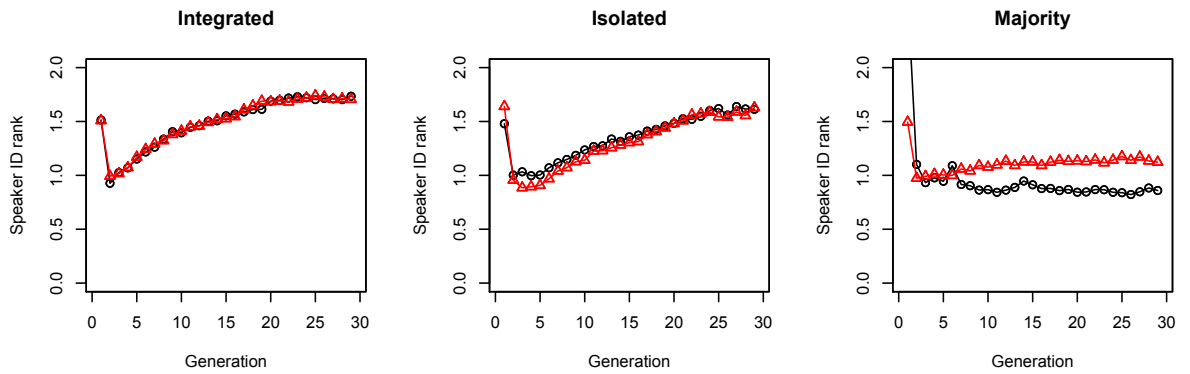


Figure 7.15: The mean relative rank of the speaker identity variable in comparison to other variables in the linguistic models of the individuals, for 100 independent simulations. The results are shown for two integrated communities (left), two isolated communities (middle) and a minority situation (right). Black lines indicate community 1 and red lines indicate community 2. In the minority situation graph, the black line is the minority community.

However, this trend is only strong in the first generation of contact (see figure C.50). During diverging generations, there is a 41% chance of a switch in superstrate community in the first two generations of divergence (from 100 simulations, significantly different from no switch: $t = 16.7377$, $df = 399$, $p\text{-value} < 0.000001$, but also random switching: $t = -3.55$, $df = 399$, $p = 0.0004$). In contact situations, there is a 49% chance change of a switch in superstrate community in the first two generations (from 100 simulations, significantly different from no switch: $t = 13.8$, $df = 199$, $p < 0.0000001$; but not significantly different from random switching: $t = -0.28$, $df = 199$, $p = 0.78$). In one generation a community might adapt to another, but this can cause the models in that community to better fit the data, leading to a pressure for the other community to adapt in the subsequent generation⁸.

7.7.2 Discussion

Comparing the results of the bottom up model with the conclusions of the top down model, four points emerge which are discussed below:

7.7.2.1 The model as a proof of concept

Previous chapters suggested that bilingualism is a population-level phenomenon and that there is no *a priori* reason to think that a human language acquisition

⁸Although a preliminary result that requires more investigation, it might be interesting to relate this to instances of ‘mixed languages’ where the emerging language in a contact situation uses the lexicon of one source language, but the grammar and morphology of the other (e.g. *Media Lengua*, see Muysken, 1997). If lexical items and morphology take different amounts of time to learn (as suggested by Clahsen et al., 2010), then the ‘mixing’ might be partially due to this alternation in the community that adapts: the lexicon is taken first from one language, and later the morphology from another.

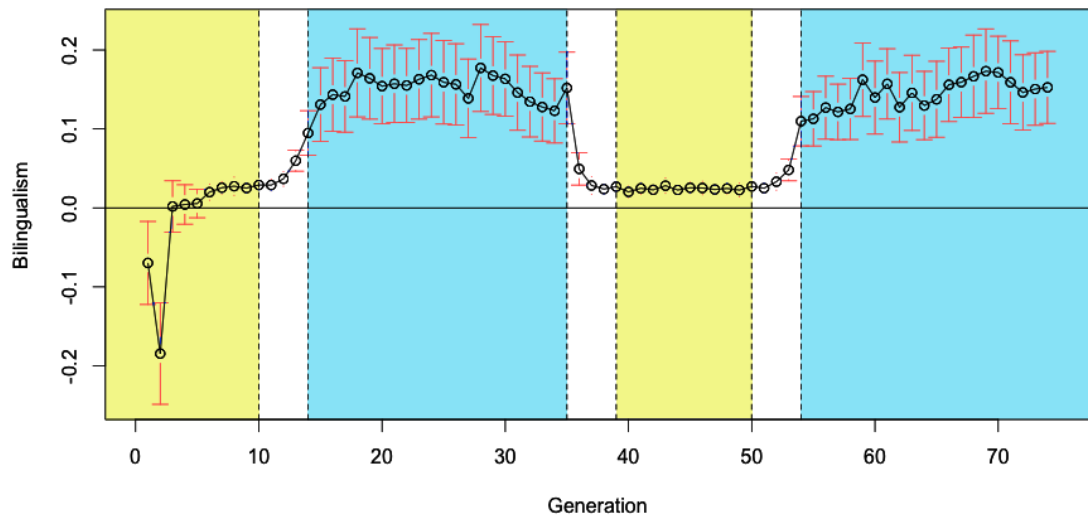


Figure 7.16: The levels of bilingualism in two communities, with two learners each, which are alternatively integrated (yellow) and semi-isolated (blue).

device has a specific mechanism for dealing with linguistic input from multiple languages, nor to think that there is a cognitive reality to discrete languages that such a mechanism could operate on. The bottom up model can be seen as a tool for thinking about this approach to language evolution. The model works as a proof-of-concept for the concrete measurement of bilingualism. The amount of bilingualism is measurable in this model without encoding a discrete, monolithic, static concept of a ‘language’. This measure behaves as we would expect. In integrated communities, it is close to zero. For two semi-isolated communities, it is negative. Furthermore, if we set up a social situation where one community is a minority, the bilingualism score captures the expected changes to the linguistic system by reflecting a positive value.

7.7.2.2 Linguistic diversity tracks social change

Dynamic social structures are a key aspect for explaining the emergence of bilingualism in this model. In the top-down model, social structures were static and so they could not form a part of the explanation. The flexibility of the bottom-up model derives from its non-rational assumptions. The results above are in line with the expectation that unconditioned linguistic variation is unstable (see Smith and Wonnacott, 2010). The linguistic contrast between communities will diminish if there is no contrast in the social variables on which the linguistic signal can be conditioned.

7.7.2.3 Addressing the bilingual paradox

Bilingualism emerges in this model without individuals having a specific mechanism for dealing with bilingualism. The top-down model specified a prior bias over the amount of variation to expect in an agent's input, fitting the learning mechanism to the problem being addressed. The bottom up model demonstrates that bilingualism can emerge from a general learning mechanism which conditions a linguistic signal on semantic variables. There are no expectations over the amount of variation to expect within or between speakers. Indeed, if social variables do not explain any of the variance, they do not play any role in an agent's internal linguistic representation. Furthermore, as mentioned above, the model maintains a division between population level phenomena and individual learning mechanisms: bilingualism can emerge at the population level without discrete, static languages being encoded in the model of individuals.

This has an impact on the bilingual paradox. Recall that this questioned why bilingualism was prevalent if learning multiple languages is difficult. The bottom up model suggests that the paradox derives from invalid assumptions by demonstrating that 'bilingualism' is a property of populations which is not necessarily related to individual learning biases. That is, whether humans have an expectation about the number of languages that will be in their input, or whether learning two languages is more difficult than learning one are not necessarily the most relevant questions to address the prevalence of bilingualism. Rather, one should ask how contrasts in social variables support the maintenance of linguistic variation. In other words, how do social structures and linguistic variation coevolve?

7.7.2.4 Implications of the model

The final point is related to the one above and concerns the kinds of questions each model is biased towards answering. The top-down model could answer questions like 'what is the rational, optimal prior expectation about variation in the input?', or in other words 'is bilingualism rational?'. It might suggest that a fruitful avenue of further investigation would be to try to estimate the bias human learners have over the number of languages to expect in their input, the amount of noise in transmission or whether the social structure was one that caused bottlenecks on learning. However, in the bottom up model, because the amount of bilingualism depends on dynamic social structures, asking whether bilingualism is the rational expectation does not make any sense without also thinking about *dynamic* social structures. This suggests that the questions asked by the top-down model are misleading. With regards to future research, the bottom up model suggests looking at dynamic social structures, and how linguistic variation is conditioned on social variables. This suggestion is explored further in the next chapter.

Both the top down and bottom up models are very abstract, and it would be

a difficult and perhaps pointless exercise to try to determine which model was more ‘realistic’ or fitted real data better. Rather, both approaches can be seen as converging on a common solution to the problem from different angles. The top down model is better at yielding good analytic results, but the bottom up model allows more flexibility in terms of social dynamics⁹. The bottom up model presented here has suggested that some of the assumptions of the top down model require more scrutiny. In response, a top down model could be built which addressed the most relevant points raised by the bottom up model perhaps using techniques such as empirical Bayesian analysis (see section 5.9.3.2 in chapter 5). Again, it should be emphasised that the difference between the top down and bottom up models provide useful insights specifically for theories of the cultural evolution of bilingualism. It may be perfectly valid to use the top down approaches discussed previously for investigating other levels of linguistic structure.

7.7.2.5 Audience design

The conclusions of the model converge on similar ideas to the concept of audience design (Bell, 1984). As mentioned above, Bell defined variation in ‘style’ as variation within a speaker that was not due to linguistic constraints. This contrasts with variation between speakers, such as differences between the speech of members of different social classes. Bell demonstrates that stylistic variation within a speaker can be affected by the identity of their interlocutor, an affect he calls ‘audience design’. The bottom up model allows both types of variation to occur, and the bilingualism measure can be seen as a measure of the amount of ‘audience design’ or ‘style shift’ in the population. The model’s conclusions are also compatible with Bell’s observations about the relationship between the two kinds of variation, which he calls the ‘axiom of style’:

“Variation on the style dimension within the speech of a single speaker derives from and echoes the variation which exists between speakers on the “social” dimension.”

(Bell, 1984, p. 151)

Bell argues that there must be variation between speakers in a community for there to be ‘style shift’ in the speech of an individual. This is exhibited in the model, as individuals only exhibit variation according to the identity of their interlocutor if there are salient differences between groups of speakers. Bell also argues that stylistic variation derives from social variation, so that social variation precedes stylistic variation. Similarly, in the model, changes to the social structure are what causes bilingualism. What the model fleshes out is the diachronic implications of this hypothesis: that bilingualism (or style shift) will reduce over

⁹Laboratory experiments with humans can be seen as a third approach which provides good realism in terms of the learning mechanism, but poor experimental control. See Scott-Phillips and Kirby (2010) and Cornish et al. (2009, p. 199) for a discussion on models and experiments being mutually supportive.

time in a population with a fixed social structure.

The concept of bilingualism used in this model is very compatible with research questions in sociolinguistics. This is because the model can track the *emergence* of linguistic categories and how they coevolve with social structure. The top down models in the previous chapters were not designed to address questions of intraspeaker variation, and so may be less compatible. In the next chapter I will argue that both the field of language evolution and bilingualism consider the co-evolution of linguistic categories and social structure, as well as learning biases. It seems to make sense, then, that the field of language evolution should be interacting with the fields of bilingualism and sociolinguistics, since they are converging on similar questions and similar conclusions from different angles.

7.7.2.6 Future research

There are many features of the model that have not been explored here, such as changing the population size, the number of communities and manipulating the semantic variable distributions. While the current analysis is enough to demonstrate the points that are relevant for this thesis, future investigations might investigate these features further.

7.8 Conclusion

This chapter presented a bottom-up model of cultural transmission which allowed bilingualism to emerge in a population. I suggested that a learning mechanism that detected conditioned variation could resolve the paradox of the emergence of a capacity for bilingualism without an apparent adaptive advantage. This general learning mechanism would be domain general. It would allow the learning of variation at multiple levels of linguistic structure by finding the most salient conditioning factors on which to structure the variation in the input.

The bottom up model contrasted with the top-down models presented in the last chapter by assuming that the linguistic signal was a continuous variable which was conditioned on semantic variables, that social structures were dynamic and that bilingualism was a property of populations which could be measured by observing how individuals condition their linguistic variation on social variables. The bottom-up model demonstrated that thinking about bilingualism in a cultural framework requires thinking about dynamic social structures. It also suggested that the bilingual paradox presented at the start of this thesis derived from invalid assumptions about bilingualism. This suggests a different perspective on bilingualism, which is further explored in the next chapter.

CHAPTER 8

THE IMPLICATIONS FOR BILINGUALISM

“The idea that people speak one language is certainly not true ... everyone grows up in a multilingual environment”

Chomsky (2000), p. 59

8.1 Introduction

The previous chapters have focussed on how a consideration of bilingualism has an impact on theories of cultural evolution. This chapter looks at how a cultural evolution approach to language can have an impact on studies of bilingualism. Bilingualism is not seen as a phenomenon that is central to language evolution or language acquisition (see De Groot, 2010 and section 8.2). Part of the problem involves the difficulty of conceptualising bilingualism: a bilingual is often thought of as being an individual who has native competence in more than one language. However, chapter 3 demonstrated problems with a categorical approach to languages, and this chapter demonstrates similar problems with categorising someone as a ‘native’ or ‘non-native’ speaker. Since a traditional definition of bilingualism rests on these concepts, it follows that there must be problems with categorising the linguistic experience of individuals into narrow categories such as ‘monolingual’ and ‘bilingual’. This chapter suggests some problems that arise from this in experimental approaches, including the exclusion of certain kinds of bilingual experience, difficulty in recruiting participants and difficulty in comparing results across studies.

A solution is offered in the form of a cultural evolution approach to bilingualism. The previous chapters have argued that languages should be treated as complex phenomena that emerge from the way individuals use language. This suggested a definition of bilingualism based on how low-level linguistic variation is conditioned on semantic and social variables. Instead of directly comparing ‘monolinguals’ and ‘bilinguals’, the cultural evolution approach suggests using an individual differences paradigm. This involves using low level measurements to see how properties of the input affect learning processes which go on to affect learners’ linguistic output. Categories like ‘monolingual’ and ‘bilingual’ can emerge from the data, but are not assumed *a priori*. Some studies of bilingualism

are already taking this kind of approach (e.g. Paradis, 2011; Parra et al., 2011).

Some implications of this cultural evolution approach are discussed. These include advantages in recruiting participants from a wider pool, the possibility of building cumulative data, clearer explanations and the ability to address practical questions about language policies. The cultural evolution approach also realises that researchers in the field of bilingualism can ask a common question: “how do linguistic variation, learning biases and social structures coevolve?”. This is a question also addressed by evolutionary linguists. I discuss tools from cultural evolution that can broaden the scope of bilingualism studies, including formal tools, computational models and experimental techniques.

Section 2 demonstrates that bilingualism has been seen as a marked condition in comparison to the ‘normal’ state of monolingualism. I argue that this should not necessarily be the case. The next few sections discuss the problems with categorising linguistic experience into narrow categories. These include categorising competence into ‘native’ versus ‘non-native’, linguistic variation into discrete ‘languages’ and language learners into ‘monolingual’ versus ‘bilingual’ categories.

Section 3 discusses the concept of native competence. Many conceptions of bilingualism involve some notion of competence. However, section 3 demonstrates that a discrete view of ‘native’ versus ‘non-native’ competence is problematic, in much the same way as chapter 3 demonstrates that the concept of discrete languages is problematic. Since this has not been dealt with explicitly yet, the issue is discussed at some length. Two opposing solutions include researchers being explicit about what they mean by ‘native competence’ and eliminating the concept of ‘native competence’ from the field of linguistics. However, such drastic measures may not be necessary, as the next sections demonstrate that researchers in bilingualism have begun to describe their work without needing to refer to concepts of ‘native competence’.

Section 4 demonstrates that experiments into bilingualism assume that languages are discrete, although there may be problems with this view. Section 5 shows that categorising linguistic experience into narrow categories leads to three problems: making unjustified assumptions about linguistic experience; difficulty in recruiting participants for experiments; and possible confounding of cause and effect in bilingual experience.

As an alternative to the categorical approach to linguistic experience, the cultural evolution approach to bilingualism is presented in section 6. This approach derives from the concept of bilingualism developed in this thesis and leads to solutions to each of the problems above. This section will also demonstrate that recent research has been moving towards this approach, circumventing the difficulties that arise from a categorical view of linguistic experience. Section 7 discusses the implications of the cultural evolution approach, and how it could

widen the scope of studies of bilingualism.

8.2 Bilingualism as the marked condition

Bilingualism is often seen as a marked case which is not a central object of study for the field of linguistics (De Groot, 2010; Cook and Newson, 2007; Gal, 2009; Sorace, 2011b). This section discusses that bias which has influenced the ‘traditional’ approach to bilingualism. As De Groot puts it:

“Several researchers have criticized the practice, common at one time, of regarding the expressions of monolinguals as the norm against which the language of bilinguals and multilinguals should be evaluated, a comparison that in the past has often led to the harsh verdict that the language use of bilinguals and multilinguals, and especially their use of non-native languages, is inferior to monolinguals’ language use.”

(De Groot, 2010, p. 340)

In fact, exposure to multiple languages is the norm in the majority of societies (Crystal and Wang, 1997; Chomsky, 2000; Wolff, 2000; European Commission, 2012). The introductory chapter mentioned some statistics: the majority of Europeans know more than one language, other countries estimate that between 66% and 80% of the population is bilingual. Despite these facts, there is a common perception, especially in industrialised western societies, that monolingualism is the norm (e.g. see Yildiz, 2012; Minnaard, 2012). Indeed, bilingualism is often approached as a marked case - the ‘default’, paradigmatic object of study being monolinguals. Chomsky’s ‘ideal speaker’ as the central object of study for linguistics clearly prioritises the study of monolinguals¹ and studies often focus on

¹Chomsky advocates the study of the ideal speaker ‘ideal speaker’ who knows their language perfectly in a homogeneous population (Chomsky, 1965). Populations where everyone speaks two languages, while homogeneous, are also excluded because “the language of such a speech community would not be “pure” in the relevant sense because it would have “contradictory” choices for certain of these options.” (Chomsky, 1986, p. 17).

Chomsky also claims that the study of a single language is sufficient to reveal general linguistic principles, suggesting that the ability to learn a single language is the most basic ability: “... In the examples I have just reviewed, I have not hesitated to propose a general principle of linguistic structure on the basis of observation of a single language. The inference is legitimate, on the assumption that humans are not specifically adapted to learn one rather than another human language. . . . Assuming that the genetically determined language faculty is a common human possession, we may conclude that a principle of language is universal if we are led to postulate it as a “precondition” for the acquisition of a single language”. (Chomsky, 1980, p. 48). Chomsky appears to claim that the above quote is “invented” (see Chomsky, 2007, p. 1096). Although this quote often appears elsewhere without the initial qualifier of “In the examples I have just reviewed” (e.g. in Evans and Levinson, 2009, p. 436; Boden, 2006, p. 645; Itkonen, 1996, p. 487), it does appear in the above citation and Chomsky re-iterates the argument (Chomsky, 2007, p. 1096).

However, Chomsky is realistic about the actual existence of these ideal speakers: “the idea

variation within a population or a language while neglecting the variation within individual speakers (Ke, 2004). The markedness of bilingualism is also evident in the language used to describe bilingualism. For instance, Roeser defines bilingualism as the “*impressive* command of two different languages” (Roeser, 1999, p.169, emphasis mine). Cook and Newson expresses surprise that “somehow two languages, two grammars can coexist within the confines of one mind.” (Cook, 2010, see also Petitto and Kovelman, 2003). In some cases, there is explicit doubt about the capacity for bilingualism, even from prominent researchers in the field of bilingualism:

“When you meet people who tell you they speak four or five languages, give them a smile to show you’re impressed, but don’t take this claim very seriously.”

Carol Myers-Scotton (quoted in Erard, 2012, p.13) ²

Petitto and Kovelman (2003) note that attitudes to bilingualism are often paradoxical: While we are impressed with the acquisition abilities of children, we are concerned that children raised bilingually will be confused or develop less quickly. This might stem from the early work on the link between bilingualism and cognitive development. Until the 1960s, studies on the link between intelligence and being raised bilingually found negative trends (see Hakuta, 1989; Butler and Hakuta, 2004 for a review). However, methodological errors were pointed out that biased results against bilinguals (Peal and Lambert, 1962). Since then, there has been a growing literature demonstrating many linguistic and non-linguistic cognitive benefits to being raised bilingually (see Bialystok et al., 2009; Lauchlan et al., 2012; Goetz, 2003)³. There is even evidence that bilingual children are less likely to develop emotional or mental health problems (Han and Huang, 2010).

The focus on monolingualism has “has arguably led to an incomplete conception, possibly even a false one, of human linguistic ability and language processing” (De Groot, 2010, p. 3). While there has recently been a lot of work on bilingualism, there is still a continuum of views about the correct approach to bilingualism. Starting at one extreme, Chomsky suggests that the object of study for linguists should be the ‘ideal speaker’ who knows their language perfectly in a homogeneous population (Chomsky, 1965 as discussed above), which some have argued

that people speak one language is certainly not true ... everyone grows up in a multilingual environment” (Chomsky, 2000, p. 59); “Bilingualism is normal to the species in the trivial sense that the world is so complex that strict monolingualism is almost unimaginable.” (Chomsky, 2000, p. 59)

²Erard (2012) discusses cases of ‘hyper-polyglots’ and their perception by academics and society. A central question of the book is how to measure a speaker’s linguistic capacities, with no formal answer, but a suggestion that the ability to successfully communicate in real, open-ended contexts with native speakers of diverse language backgrounds is an index.

³Many domains have been slow to internalise this, however, a recent case being an expert witness testifying in a court of law that Welsh-medium education in Wales could cause ‘mental retardation’ (Shipton, 2008).

has biased researchers towards only looking at monolinguals (Sober, 1980; Cook and Newson, 2007)⁴. Grosjean argues that researchers cannot compare monolinguals and bilinguals directly (Grosjean, 1985, 1989). Cook (1995) suggests that researchers should *only* study bilinguals, and others such as Sternberg and Christiansen (2006) suggest the bilingual ability is a central part of the faculty of language. Some claim that there is no such thing as a pure monolingual (Roeper, 1999; Otsuji and Pennycook, 2010; Pennycook, 2010; De Groot, 2010).

I argue, given the approach of this thesis, that the dichotomy between monolingual and bilingual is misleading. Chapter 3 showed that there were problems with a discrete view of languages, and so the conception of bilinguals as individuals with knowledge of two is problematic. The measure of bilingualism suggested in this thesis sees bilingualism as a continuous measure of the amount of linguistic variation that is conditioned on social variables. As part of this argument, in the next section I will demonstrate that a central criterion for identifying bilinguals - native competence - is also a problematic measure.

8.3 Categorical approaches to native competence

Definitions of bilingualism often involve concepts of native competence (see section 1.2.1 in chapter 1). However, the problem of classifying speakers as ‘native’ or ‘non-native’ is problematic theoretically (e.g. Escudero and Sharwood Smith, 2001; Davies, 2003; Medgyes, 1992; Hyltenstam and Abrahamsson, 2003) and empirically (e.g. Dabrowska, 2010; Street and Dabrowska, 2010; Mulder and Hulstijn, 2011.). There certainly are maturational effects on language acquisition abilities (see Long, 1990; Birdsong and Molis, 2001 for reviews). However, labelling this difference with a discrete category may be problematic because the concept brings with it implications about various other properties that can affect choices made in the design of experiments.

In this section I discuss these problems and three solutions. First, researchers can continue using the concept as long as they explicitly state their definition of native competence (Escudero and Sharwood Smith, 2001). However, this renders the concept of a ‘native speaker’ of little value, so a second solution is to eliminate the concept from the field. That is, researchers should stop using a categorical distinction between ‘native’ and ‘non-native’ speakers. This section

⁴“This emphasis on monolingualism has simply been taken for granted by those working within the UG theory, along with the other areas excluded from competence, and is seldom discussed or justified. The only true knowledge of the language is taken to be that of the adult monolingual native speaker.” (Cook and Newson, 2007)

“Additionally, Chomsky remarks that the properties of such social entities are ‘artefacts’ or are ‘epiphenomenal’, meaning, presumably, that they are just the causal upshots of interactions at the level of individual psychology. So even if line-drawing problems could be solved, there still would not be a reason for thinking that socially shared languages are needed to explain anything.” (Sober, 1980, p. 379)

is mainly concerned with setting out a formal argument for elimination. However, it is conceded that outright elimination may be difficult to achieve because the concept of a native speaker is entrenched in the field of linguistics. Instead, there may be a third alternative between keeping the categorical concept of native competence and actively eliminating it. The next sections demonstrate that researchers in bilingualism have started to avoid classifying experimental participants into discrete categories such as ‘native’ and ‘non-native’ or ‘monolingual’ and ‘bilingual’. Instead, they use low-level properties as direct measures in their experiments, bypassing the need to define what ‘native competence’ refers to.

8.3.1 Native competence as a prototype

Escudero and Sharwood Smith (2001) suggest the following definition as characterising a restrictive definition of a native speaker:

“A native speaker of *Di* is someone who grew up in a community of speakers where (i) only *Di* was spoken, and (ii) the linguistic behaviour of the individual in question is perceived both by members of that community of speakers, and by the individual him/herself, to be that of a full member.”

(Escudero and Sharwood Smith, 2001, p. 277)

Anyone whose linguistic performance matches this group should qualify as native. The definition makes three assumptions. First, one can identify a single language. Secondly one can identify a discreet community. Finally, that the perception of linguistic identity is straightforward. Chapter 3 argued against these assumptions and demonstrated that the concept of a ‘language’ is related to many linguistic and non-linguistic factors. Similarly, Escudero and Sharwood Smith (2001) identify a set of prototypes that are typically thought of when classifying a speaker as native. These include a set of core prototypes which most linguists would agree as being central to the concept of native competence, such as knowledge of the lexicon and grammar, accent, the initial language environment (i.e. if there was exposure from birth) and whether language use was maintained after initial exposure. There are also more marginal prototypes such as fluency, literacy, pragmatics and grammatical intuitions. Finally, there are peripheral prototypes such as socio-cultural knowledge, paralinguistic knowledge, identity and orthography.

A researcher’s definition of a native speaker typically uses these prototypes. For instance, a native speaker of a language is an individual who was exposed to the language from birth. While Escudero and Sharwood Smith admit that there is little consensus about which prototypes are most important, they suggest that as long as authors are explicit about what they mean when they refer to native competence, then it can be a productive concept. An alternative to this would be to eliminate the concept. An argument for doing this is set out below.

8.3.2 Eliminating the native speaker

As the section above suggested, there are many properties that can go into classifying a speaker as having native competence. While there may be prototypical ideas of the properties of native speakers, they are not a homogeneous population. For example, experimental work has shown that adults who fit typical prototypes of native speakers can have different grammatical intuitions, according to their level of education (e.g. Dabrowska, 2010; Street and Dabrowska, 2010; Mulder and Hulstijn, 2011). Furthermore, the competence of speakers can undergo attrition over time if the speaker does not maintain use of a language (Lambert and Freed, 1982; Seliger and Vago, 1991; Andersen, 1982; Sorace, 2004). If a ‘native speaker’ does not refer to a stable concept, then perhaps it is not useful or even misleading.

Based on similar observations, section 3.13 in chapter 3 suggested that the concept of languages as discrete, monolithic, static entities could be eliminated. Here I apply the same methodology to suggest why the concept of native competence is also a candidate for elimination. Essentially, if a ‘native speaker’ picks out learners by a single property, for instance age of first acquisition, then there’s little point having a separate term for it: we can just use ‘age of first acquisition’. If it is intended to also include frequency of exposure, competence and attrition, then it seems like a confused concept that is better thought of as a set of separate measures and mechanisms, each of which can be studied separately. Similar criticisms of other fields have suggested that concepts such as ‘consciousness’, ‘emotions’ and ‘concept’ can be productively eliminated from a given field (Irvine, 2011; Griffiths, 1997; Machery, 2009).

8.3.2.1 Identifying the target phenomena

If the concept that ‘native competence’ refers to is not coherent, then the concept may be productively eliminated from the field. One way to demonstrate this is if the criteria for the judgement of the concept is sensitive to the context of a study (see Irvine, 2011). Non-linguistic factors can affect people’s perception of the native competence of a speaker in an effect called ‘reverse linguistic stereotyping’ (Rubin et al., 1997; Kang and Rubin, 2009; Rubin, 2011). In a series of experiments, English-speaking participants listened to speech from a native speaker of English. However, in one condition the speaker was presented visually as having a non-local cultural identity. Participants rated the same speech audio as having a less ‘native’ accent in this condition. The overall comprehension of the speech was actually worse in this condition. This suggests that judgements of ‘native’ competence can be manipulated by non-linguistic factors and the context of the task. This means that the concept of a native speaker may not refer to a coherent phenomenon. Further problems with the concept are discussed below.

8.3.2.2 Argument from the scientific method

The scientific method usually proceeds in the following stages. Step 1: discover a dissociation. Step 2: Categorise the phenomena. Step 3: test the categorisation. A final step may involve discovering problems with the categorisation, and so seeking more precise dissociations (back to step 1). However, some experimental studies in linguistics (see example below) have first assumed that there are differences between ‘native’ and ‘non-native’ speakers (step 2) and then tried to find measures by which they differ (step 1). These measures are usually motivated by intuitions about how ‘native’ and ‘non-native’ speakers differ. This goes against the usual order of the scientific method.

For example, Johnson and Newport (1989) tested the grammatical knowledge of English of immigrants to the United States of America. They found that the age of arrival was the only significant predictor of attainment. To test the critical period hypothesis, Johnson and Newport split the groups into those who arrived before the age of 16 and those who arrived afterwards. For the early learner group, they found a significant negative correlation between age of arrival and attainment (the earlier the exposure to English, the higher the ultimate attainment is). However, there was no significant correlation for the late learner group. Johnson and Newport claimed this was evidence for a critical period that constrains the ultimate attainment possible in a second language.

However, Birdsong and Molis (2001) replicate this study with highly proficient Italian learners of English and find the opposite effect. There was a significant correlation for the late learners, but not for early learners. Therefore, the question of whether adult L2 learners can achieve native proficiency is sensitive to the selection criteria for experimental participants (see Hyltenstam and Abrahamsson, 2003 for a review).

Hyltenstam and Abrahamsson (2003) point out that part of the source of the conflict is that the studies split the participants into two groups, then test the correlation between age of first exposure and ultimate attainment. That is, they were assuming a dissociation after a cut-off point, then trying to measure this dissociation. In both studies, taking the results for all ages together, there is a gradual decline in ultimate attainment with age.

However, there are learners who were exposed to a language after puberty who have an ultimate attainment within the range of native speakers (van Wuijtswinkel, 1994; Cranshaw, 1998; Bongaerts, 1999; White and Genesee, 1996; Sorace, 1993; Sorace et al., 2009, see Birdsong and Molis, 2001). Hyltenstam and Abrahamsson conclude that “it is inherently difficult, perhaps even impossible, to distinguish native from near-native speakers. The slight differences that exist between them may well be unnoticeable” (Hyltenstam and Abrahamsson, 2003, p.571). They suggest that a resolution of the literature on matura-

tional constraints on ultimate attainment requires recognising a category of “non-perceivable non-nativeness”. However, this appears to mean that these speakers are only classified as ‘non-native’ based on their age of initial exposure. That is, “non-perceivable non-nativeness” is not a measure of language attainment, but of life history. If this is the case, then the usefulness of a discrete division between ‘native’ and ‘non-native’ speakers in an explanation of maturational constraints is unclear. The scientific approach allows the concept to be eliminated if the dissociations turn out to be a collection of finer-grained dissociations (e.g. the ‘critical period’ effect is a product of maturational effects, age of exposure, exercise and social and psychological effects, as suggested by Hyltenstam and Abrahamsson, 2003). However, assuming the concept of the ‘native speaker’ is real before finding a real dissociation can make the concept of the ‘native speaker’ resistant to development.

8.3.2.3 Underspecification

As we have seen above, there are many definitions of the concept of a ‘native speaker’ such as age of acquisition (both using fine-grained and cut-off measures, e.g. puberty), competence and frequency of exposure. However, there are other concerns that might affect the competence of a speaker such as possible differences in comprehension versus production (Lincoln, 1979; Croft, 2003), the distance between the L1 and L2 (Lado, 1957), ‘native’ accent (Kang and Rubin, 2009; Rubin, 2011) and the level of attrition (e.g. Andersen, 1982; Seliger and Vago, 1991; Sorace, 2004). In a minority scenario, motivation to learn and use a language and socio-political factors may have an effect on the competence (Krashen, 1982). This plethora of factors suggests that the concept of a ‘native speaker’ is under-specified the phenomenon it tries to describe.

8.3.2.4 Taxonomic errors

A speaker may be ‘non-native’ according to a range of mechanisms. For instance, if the speaker is very young (development) or the speaker has not used the language in a long time (attrition). Lumping together the results of different mechanisms under one concept can be confusing. Also, as we have seen above, the concept of speakers who are indistinguishable in terms of competence may be labelled ‘native’ or ‘non-native’ based on their life history. Therefore, the concept also divides individuals that are very similar. Furthermore, different criteria may be applied when assessing nativeness according to other properties of the speaker. Adults who achieve native competence in a second language are “the object of much admiration and astonishment. For child learners, however, everything short of nativelike levels is seen as a failure.” (Hyltenstam and Abrahamsson, 2003, p. 539).

These three points suggest that the taxonomy of native competence is confused. As we shall see in the rest of this chapter, recent experimental work in bilingualism has been ignoring the categories of ‘native’ and ‘non-native’ and instead

talking about specific mechanisms directly.

8.3.2.5 Misidentification

The sections above have shown that there are a range of ways to operationalise the concept of a ‘native speaker’. As research reveals the increasingly complex mechanisms behind acquiring a linguistic system that aligns with a population of speakers, the usefulness of a black-and-white concept like ‘native’ and ‘non-native’ can be questioned. At the very least, the use of the concept requires clarification in terms of the features that are being referred to (Escudero and Sharwood Smith, 2001). The rest of this chapter will argue that explaining research in terms of specific mechanisms is more productive than using categorical labels. However, it is quite another question whether the concept should be actively eliminated. Below are some arguments for keeping the concept.

It was then suggested that the concept has been made redundant as the field has progressed to consider finer-grained measures of native competence such as age of acquisition, frequency of exposure, attrition and so on.

8.3.3 Arguments against eliminating native competence

The sections above argued for the elimination of the categorical distinction between ‘native’ and ‘non-native’ speakers. However, outright elimination might be difficult to achieve for four reasons, detailed below.

First, the concept of the ‘native speaker’ may be a useful generalisation. The vast majority of formalist approaches to language depend on a ‘native speaker’ assumption. Secondly, the concept is entrenched. The ‘native speaker’ is a concept that is the very cornerstone of generative linguistics. However, in the experimental community, researchers are moving towards finer-grained measures.

Thirdly, the concept may promote stability. The very question of how children achieve ‘native’ competence is the motivation for many studies of second language acquisition. Eliminating the concept could cause confusion in the field.

Fourthly, the concept may be used unambiguously. ‘Native speaker’ is usually shorthand for ‘learned this language from birth’ as opposed to ‘learned as an adult’. While there is little consensus about what is meant by a ‘native speaker’ between studies, the concept may be clear within studies (Escudero and Sharwood Smith, 2001). If the concept can be used unambiguously, it may not be problematic.

Despite these arguments, a categorical concept of ‘native’ versus ‘non-native’ could still be problematic because it can affect the choices made in experiments. For example, many linguistic and psycholinguistic studies require ‘native speak-

ers' as a control. Studies will often use 'native' participants who self-identify as 'native', without applying scientific measures to check their actual competence (see De Groot, 2010, p. 3). In contrast, many measures are often applied to individuals who are considered to be 'non-native'. However, there may be a middle ground between eliminativism and maintenance, discussed below.

8.3.4 Summary

The sections above argued that the concept of a discreet difference between 'native' and 'non-native' speakers may not be useful for explaining phenomena in bilingualism. It is difficult to apply the concept consistently while still being more useful than appealing to lower-level features directly. Despite this, outright elimination may be counter-productive because the concept is entrenched and historically important. However, the rest of this chapter demonstrates that research is being carried out in the field of bilingualism that does not rely on the concept of 'native speaker'. Perhaps, then, as research reveals the intricacies of mechanisms that affect language learning, the high-level category of 'native' versus 'non-native' will be naturally replaced with lower-level explanations.

8.4 Categorical approach to languages

Chapter 3 discussed the difficulty of counting the number of 'languages' that an individual speaks. This poses a problem for studies of bilingualism since bilinguals are usually identified using this metric. This problem is discussed in chapter 3. An additional problem with a categorical approach to languages is that it also affects choices about experimental conditions and stimuli, discussed here. For example, participants may be primed with stimuli that are assumed to activate a particular language 'mode'. However, even with what are traditionally thought of as well-defined languages, categorisations of linguistic variants can be very subjective. I will illustrate this with an example from Thomas and Allport (2000), but this problem applies to many studies.

Thomas and Allport (2000) conduct a study of task switching in French-English bilinguals which used words as stimuli. They state that "no words were used that existed in both languages" (Thomas and Allport, 2000, p.47). However, the truth of this assumption depends on the definition of how to categorise words into languages. We can test this statement by assuming that the words categorised as French will not appear in English usage. However, 22% of their 'French' stimuli appear in the 'English-language' Brown corpus. For example, 'bureau' was assumed to be an exclusively French word, but is within the 2,500 most common words in the English corpus.

Although Thomas and Allport begin with this subjective definition, they actually then proceed to use objective measures to show that the subjective categorisa-

tion is valid. This is done in two ways: First, they asked bilinguals to rate how unique words were to English or French in terms of their spellings (the experiment involved orthographic stimuli). Secondly, they conduct an analysis of trigrams: “To verify that the stimulus sets contained, respectively, shared and language-specific orthography, the candidate sets were subjected to an analysis of their constituent, overlapping letter trigrams in terms of their frequency of occurrence in each language. The more language-specific the orthography, the less probable the occurrence of (at least) some component letter trigrams in the “other” language” (Thomas and Allport, 2000, p. 47).

Although both of these measures are more objective, they suffer from a second-order problem, since the criteria depend on how words are categorised into languages. Rather than try to find a better measure of how to categorise languages, the difficulties pointed out above might indicate that the experiment is getting things back-to-front. The usual scientific method is (step 1) to discover a measurement that shows a dissociation (e.g. switch costs) and then (step 2) label that dissociation (e.g. different ‘languages’). A characterisation of the experiment above (and many others) is that a dissociation is assumed to exist between languages (step 2), and then the experiment demonstrates that this dissociation can be measured (step 1).

Despite these problems, for the purposes of Thomas and Allport’s paper, their assumptions are reasonably supported by their measures. However, they ignore these measures in the final analysis of the experimental results. The final analysis is based on their original categorical, subjective measure rather than the more fine-grained measures based on individual ratings and trigram overlap. If the objective, low-level measures are there to be used, then they should be. For example, they could ask whether the words that clustered according to the ratings and trigram measures also clustered according to switch latency times. Although the statistics may be more complicated, there are a range of new methods that help with this type of analysis, for example mixed effects models (e.g. Baayen et al., 2008; Jaeger, 2008).

Another way of priming a language mode is to do it explicitly. However there are different ways of doing this. Au and Glusman (1990) conduct experiments into children’s mutual exclusivity behaviour. In one experiment, monolingual children were presented with several novel objects and an experimenter labelled one novel object with a novel English word. Another experimenter then taught the child a novel label for an unspecified object in Spanish, saying “Do you speak Spanish? Do you want to learn a word in Spanish? Okay, now I am going to teach you a word in Spanish. *Theri* is a Spanish word. Can you say *theri*? Well, *theri* is the Spanish word for a kind of animal which is here” (Au and Glusman, 1990, p. 1487). As is evident, this explicitly presents the two languages as discrete, monolithic entities that are either possessed or not. Participants were tested to see if they would be happy to assign two labels to an object if the labels came

from different languages. Participants were asked “Can you guess which one Spanish-speaking children would call a *theri*?”. Merriman and Kutlesic (1993) argue that these instructions could bias the child’s decision “not only by reminding the children of the source language of the test word, but also by focusing their judgment on how a group that did not know the English label would use the name.” (Merriman and Kutlesic, 1993, p. 246-247).

As an alternative methodology, Merriman and Kutlesic present the children with a doll that use the novel language in a naturalistic situation and are then asked “help Suzi pick out the [*theri*] in the store.” Merriman and Kutlesic argue that

The instructions in the current study focused children’s judgement on persons who knew the English label: namely, themselves and the doll they were supposed to help. Moreover, monolinguals who are asked to think about how names are used by different groups may be prompted to reflect on the social conventional nature of naming (i.e., on how names are not inherent in objects, but are assigned to them by group consensus) and therefore may realize that an object can be assigned more than one name.”

(Merriman and Kutlesic, 1993, p. 246-247)

The approach of Merriman and Kutlesic above is the one that makes sense given the approach of this thesis. Instead of assuming that children are aware of categorical languages that align with the experimenters’ intuitions, children can be presented with characters that demonstrate a particular language mode through use.

8.5 Categorical approach to bilingualism

The sections above argued that the categorical approaches to native competence and boundaries between languages are problematic. Despite this, these concepts are often used to divide experimental participants into ‘monolingual’ and ‘bilingual’ groups. Although this problem is recognised in the literature, it is usually addressed by checking that the participants conform to a particular prototype of a monolingual or bilingual, similar to the suggestion of Escudero and Sharwood Smith (2001), criticised above. This causes three problems, discussed below.

8.5.1 Problem: Assumptions about linguistic experience

Firstly, it makes possibly invalid assumptions about the linguistic experience of both groups. Monolinguals are often assumed to be “lacking any knowledge of any language other than their native language and who do not differ from the bilinguals in other respects such as socioeconomic status, education level, cultural

background, age and intelligence”, which might be unrealistic (De Groot, 2010, p. 342). Many linguistic and psycholinguistic studies require ‘native speakers’ as a control, but often recruit these individuals through self-selection. While many measures are often applied to individuals who are considered to be ‘non-native’, these measures are often not applied to those considered to be ‘native’ (see De Groot, 2010, p. 3).

One way to control for these assumptions is to only test participants who fit a narrow definition of bilingualism. This may mean that only certain types of bilingual experience are scrutinised, possibly overlooking crucial data (Byers-Heinlein and Werker, 2009b; Sorace, 2011b). A narrow view of bilingualism can also constrict the pool of possible participants by excluding participants who do not fit a prototype.

Furthermore, assigning participants to categories suggested by a theory can lead to circularity. For instance, groups that may be defined as ‘monolinguals’ and ‘bilinguals’ may have different executive control profiles. However, if the meaning of ‘bilingual’ reduces to “people with experience of increased linguistic variation” and better executive control means “better at handling increased variation”, the result is not so surprising. It might turn out that the ‘bilingual’ group tends to experience a greater range of non-linguistic cultural practices, or a larger number of speakers, and that this is a better predictor of executive control ability. This hypothesis is addressed in a pilot study presented in section 8.7.2.1.

8.5.2 Problem: Recruiting participants

One of the main difficulties of traditional bilingualism experiments is finding participants that fit the experimental criteria. This has led to the strange situation where researchers often gather a lot of low-level, detailed information about the linguistic and cultural background of their participants in order to check that they meet the criteria, but not using these measures in their studies. The following quote describes a fairly typical set of criteria and considerations for participants in Serratrice et al. (2011):

“The following selection criteria were applied: no history of language impairment or hearing loss; bilingual children were included in the study only if they had been regularly exposed to both languages from birth and used them on a daily basis with similar competence according to teachers’ assessment and parental reports. ... The children selected for the study were growing up in households where both languages were spoken by both parents; the majority of the parents only used their mother tongue with the children, and the children generally matched the parent’s language choice when talking to them. The bilingual children in Italy had exposure to English at home through one of their parents, at school through the curriculum that was taught predominantly in English, and during visits to the United Kingdom or the United States. ... In contrast, the bilinguals in the United

Kingdom had access to Italian in the home through one of the parents, through other bilingual English-Italian children in some limited way, and through visits to Italy during school holidays. ... The monolingual control groups ... had no functional competence in a second language, although most of the Italian speakers had received some formal language instruction in English at secondary school level.”

(Serratrice et al., 2011, p. 11)

There are a lot of considerations here of both linguistic, cultural and social variables. As must be obvious, finding participants is difficult, time-consuming and expensive. Many potential participants have to be excluded. Furthermore, having collected all this detailed data on the cultural and social background of individuals, none of this was used in the main statistical analyses.

In another example, Treccani et al. (2009) conduct an experiment on spatial priming in monolinguals and bilinguals. They collect the following data for each participant: a language proficiency test, a nonverbal reasoning test, years of education, handedness, gender, proficiency in other languages, countries lived in, age of first exposure, use of language on a daily basis throughout their life, current average use of each language and self-reported fluency in each language. However, none of this data is used in the main analysis. Instead, a sub-set of the data is used to demonstrate that monolingual and bilingual groups do not differ in their averages, meaning that these factors are controlled for. That is, it is reasonable to assume that the monolingual and bilingual groups are not different except for the number of languages they have knowledge of. This means that not only do participants have to conform to the criteria, but the two experimental groups also have to match. Researchers recognise that ideal bilingual criteria may not always be possible:

“Although adopting such stringent criteria may yield an elegant research design, implementation will quickly run afoul of reality. Very few children have balanced input, and the balance of language exposure changes as family circumstances and compositions change. We have taken the approach of trying to capture the variability in children’s bilingual environments (and then studying its consequences) rather than trying to restrict the variability through selection criteria.”

(Hoff, 2011, p. 317)

This approach is similar to the individual differences solution discussed below. However, Hoff still uses a categorical approach to labelling participants, claiming that “even with this inclusive approach, we have to reject some potential participants because they cannot clearly be categorized as bilingual or monolingual by our criteria” (Hoff, 2011, p. 317).

8.5.3 Problem: Confounding mechanisms

Another problem with categorising participants into ‘monolingual’ and ‘bilingual’ is that it can confound the causes and effects of bilingualism. For example, Treccani et al. (2009) begin by asking “whether the continuous experience of handling two languages, and the mechanisms that bilinguals develop to control the two language systems, have any repercussions on nonlinguistic cognitive abilities” (Treccani et al., 2009, p. 320). They conclude that “our results represent evidence in support of a substantial effect of bilingualism on executive control processing” (Treccani et al., 2009, p. 326). However, there is a disconnect between the question and answer. They assume that one of the mechanisms that cause the differences in the participants so that they come to be categorised as ‘bilingual’ (e.g. handling different languages) is the same one that causes the difference in the executive control tasks. While there is good reason to believe this is true, the research question could be addressed more directly. One prediction is that bilinguals who use each language equally on a day-to-day basis would have better executive control than those who only used one language occasionally. In fact, this question could be addressed with the existing data collected by Treccani et al..

Some current research questions involve complex relations between variation in linguistic input, social structure, development, linguistic output and non-linguistic abilities. For instance, to what extent does dealing with linguistic variation improve executive control in non-linguistic domains (see Calabria et al., 2011, see section 5.9.3.1)? There’s a danger that categories like ‘monolingual’ and ‘bilingual’ which are effects of lower-level experience could be reinterpreted as causes of effects in other domains. This might not be valid, since fitting linguistic experience to these categories involves simplifying assumptions. The effects of these assumptions could be amplified when extending them into complex relationships like the one mentioned above, or into evolutionary models where the output of one generation becomes the input of the next generation.

As the next section will demonstrate, a cultural evolution approach to bilingualism allows low-level properties in the input to be linked directly to low-level properties in the output. This avoids having to fit linguistic experience into narrow categories in the first place.

8.6 A cultural evolution approach to bilingualism

The sections above presented some problems with using categorical approaches to bilingualism. As a potential solution, this section introduces the idea of the cultural evolution approach to bilingualism, suggested by the argument of thesis and the bottom-up model. Central to this approach is to realise that although

there are differences in cognition between ‘monolinguals’ and ‘bilinguals’, these individuals begin life with a common cognitive profile. This common profile is then affected by linguistic input and social factors that shape what is learned, but also how it is learned. That is, there is a co-evolution of input and learning biases that deal with the input.

Taking one step further leads to a cultural evolutionary perspective: The above process affects the linguistic output of a learner which goes on to be part of the input for, and shape the learning processes of, another learner. This is the iterated learning model conception of cultural evolution (Kirby, 2001; Smith et al., 2003; Griffiths and Kalish, 2007; Kirby et al., 2008)⁵. There are therefore two levels of adaptation. Firstly, the learning biases of an individual adapt to the linguistic input within an individual’s lifetime. Secondly, the language is adapting to the learning biases of individuals over cultural time (many generations). Therefore, features such as the distribution of variation in a population (the levels of ‘bilingualism’) are part of a complex adaptive system where learning biases and the input to be learned are co-evolving.

Studies of cultural evolution have developed theories of how language adapts to learning biases (e.g. Kirby, 2001; Kirby et al., 2008; Smith et al., 2003), but rarely the converse (although see Smith and Thompson, 2012, discussed in section 5.7). It is possible to see the field of bilingualism as filling this gap. The cultural evolution approach to bilingualism suggests that the main research question for bilingualism should be how linguistic variation, learning biases and social structures coevolve. An example would be the effect of being exposed to socially stratified linguistic variation on executive control (e.g. Bialystok and Craik, 2010; Hernández et al., 2010; Bialystok and Martin, 2004), which in turn supports interactions in a society where linguistic variation is socially stratified.

Conceptualising the input and output as mediated by a learning mechanism is similar in many ways to the Language Acquisition Device (LAD) conceptualisation of language learning (Chomsky, 1965). The LAD conceptualisation sees linguistic input being processed by general learning principles yielding an output of the grammatical knowledge of a language, which produces linguistic output. However, the cultural evolution approach to bilingualism differs in two ways⁶. First, a minor point, it recognises that the linguistic input was itself the output

⁵This argument is similar to that of Chater and Christiansen (2010) who argue that language acquisition should be looked at in the light of language evolution. They argue that language acquisition does not need to be viewed as a hard task, because culturally transmitted systems will be adapted towards the child’s learning biases. Therefore, a learner’s intuitions are likely to be correct. However, while this may be true for acquiring a single language, acquiring multiple languages with possibly conflicting features may be hard if we imagine the learner as having simple biases. In order to be able to co-ordinate with multiple linguistic systems, a learner would at least have to have flexible biases, at which point the question of whether learning is easy becomes more complicated.

⁶Although the original conception of the LAD is by now somewhat of a straw man, the comparison is useful.

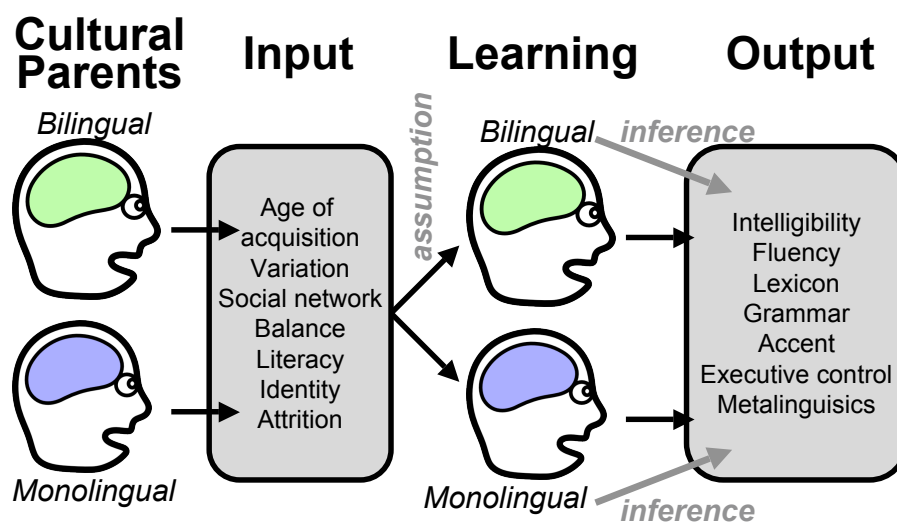


Figure 8.1: A diagram of the traditional approach to studying bilingualism. Participants are assumed to belong to either a bilingual or monolingual category, according to features of the input (which also depend on the category of their cultural parents). Inference about processes of language learning is made between these categories and properties of the output.

of a previous learner (as in the iterated learning model, see Smith et al., 2003). Secondly, the cultural evolution approach to bilingualism sees the states as concurrent rather than sequential: The ‘grammatical knowledge’ can change over time (Andersen, 1982; Seliger and Vago, 1991; Sorace, 2004), and continuously interacts with aspects of processing (Sorace, 2011b, 2004). Finally, the general learning principles are general, flexible biases that can adapt to the input (see section 5.9.3 on page 82).

The cultural evolution approach to bilingualism naturally fits with an individual differences approach to experimental design (e.g. see Skehan, 1989; Bates et al., 1991; Skehan, 1989; Vogel and Awh, 2008; Kanai and Rees, 2011). Studies that take account of individual differences use low-level measurements of multiple factors as independent variables, then use statistics to explore which factors of the learner’s input affect specific aspects of their output. Rather than try to define bilingualism more precisely, individual differences studies avoid putting participants into binary categories. The individual differences approach has several benefits. These include broadening the pool of possible experimental participants, allowing studies to use each other’s results, prioritising explanations in terms of specific mechanisms and answering questions that are important to bilingual families. These are discussed below.

8.6.1 Solution: Controlling for linguistic experience

Rather than seek groups that are not different except for the measure that is the focus of the experiment, an individual differences approach takes advantage of the

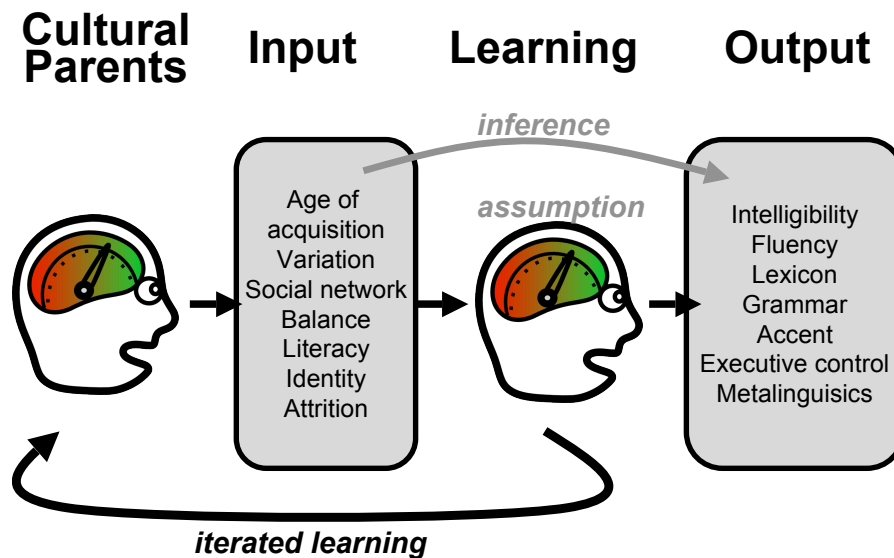


Figure 8.2: A diagram of the individual differences approach to studying bilingualism. Learners have flexible biases that are affected by features of the input. Inference takes place between features of the input and features of the output directly.

variation in the participants. Under this approach, participants are not split into two groups, but many low-level measurements are modelled statistically to determine which of them predict the performance in the experimental task. Some studies use statistical methods such as linear mixed effects modelling (Baayen et al., 2008; Jaeger, 2008) to achieve this (e.g. Ota et al., 2010; Hatzidaki et al., 2011). The differences between ‘monolingual’ and ‘bilingual’ individuals should emerge from the data. If the researchers feel that participants do clearly belong to a ‘monolingual’ and ‘bilingual’ group, then this can be introduced as another factor.

Recently there have been examples of research that move away from questions such as ‘What is the difference between monolinguals and bilinguals?’ towards questions that address individual differences. For example, Paradis (2011) conducted an extensive survey of English acquisition by children from newcomer families to Canada. Rather than compare these children to ‘native’ child learners, Paradis took a range of measures of input and competence factors in order to ask whether child-internal or external factors had the largest affect on levels of attainment. There were two types of input factors. The child-internal factors included the following:

- Age at testing
- Age at onset of English
- Comprehensive Test of Phonological Processing (phonological short-term memory measured by digit span and non-word repetition)

- Columbia Mental Maturity Scales (non-verbal IQ scores as a measure of analytic reasoning)
- Whether the L1 marked tense and agreement on verbs

The child-external factors included the following:

- Months of exposure to English
- Proportion of English spoken among family members in the home
- Number of older siblings
- Mother’s self-rated fluency in English
- Mother’s education in years
- Richness of the English environment outside school
- The child’s experiences with media, organized activities and playmates

The measures of competence included the following:

- Receptive vocabulary (Peabody Picture Vocabulary Test)
- Verb morphology test (Test of Early Grammatical Impairment)

The range of measures employed in this study are indicative of a shift away from studying differences between two groups towards studying how specific aspects of the input affect specific aspects of linguistic performance and competence. Statistical analyses were used to explore this question. This involved mainly linear regressions. Given the amount of detail in the data, more powerful statistical techniques such as linear mixed effects modelling, which takes advantage of individual differences, might have been more insightful⁷. Nevertheless, studies like this one are sure to be productive in the future.

A similarly detailed questionnaire was used in Unsworth et al. (in press). This included details about individuals interacting with the child, the amount of time the child spends in different contexts and doing different activities and the proportions of languages used in each context or during each kind of activity. As well as traditional measures such as length of exposure, age of onset and age of testing, a measure of cumulative exposure was used (the ‘Utrecht bilingual language exposure calculator’ see Unsworth, in press). This measure involves collecting data about the proportion of a child’s input in a given language at regular periods of their life. The cumulative exposure is then the number of years they have been exposed to the language modulated by the proportion of exposure to that language. This avoids confounding age of onset with amount of exposure. For

⁷Indeed, Paradis speculates “Perhaps the role of maternal education is modulated by interactions with other factors.” (Paradis, 2011, p. 232), and has the data that could test this, but does not attempt an analysis.

instance, a child can be exposed from birth to two languages, but receive very different proportions of input from each language. The measure is a useful, low-level factor that helps explain variation in language acquisition data (see Unsworth; Unsworth et al.). However, a confound that persists in the cumulative measure is the possible change in levels of exposure to different languages over a child's life. Either this could be included as a separate measure (i.e. the standard deviation of proportions from year to year), or the statistical analyses could use the year-by-year proportions directly. A mixed effects modelling approach could test whether change over time was an important factor to include in the main analysis.

In another example, Parra et al. (2011) investigate the extent to which phonological memory (storing sequences of sounds) is shaped by language experience. The participants were children raised with exposure to Spanish and English. Their relative exposure to each language, their vocabulary size in each language and their grammatical complexity were compared with their phonological memory for English-like non-words and Spanish-like non-words. Hierarchical regressions were used to demonstrate the extent to which their language exposure predicted their phonological memory

The results showed the same links between phonological memory and language development as found in monolinguals (Gathercole, 2006), but also found effects of the proportion of exposure to each language. Parra et al. suggest that language exposure supports the development of language-specific phonological memory skills that, in turn, support language development. The conclusions of the research address the research question directly and are stated directly in terms of the direct measures involved: Parra et al. state that "language exposure showed language-specific relations to phonological memory and to language development and that phonological memory partially mediated the effect of exposure on development" (Parra et al., 2011, p. 124).

There were no monolingual controls, and the focus of the paper is not the difference between monolinguals and bilinguals, but the insight into the development of learning biases. That is, effectively using bilinguals as their own control, an idea that is gaining currency:

"Using other bilingual speakers as a term of comparison, rather than only the 'classic' monolingual speaker, is not only methodologically more sound (as has been known for a long time by researchers working on child L2 acquisition - see Schwartz, 1998) but helps to see the 'forest' of a general model of bilingualism beyond the individual 'trees' of bilingual types. This in turn is a concrete move away from the concept of bilingualism as the 'sum of two monolinguals'."

(Sorace, 2011b, p. 27)

In the traditional approach, despite researchers' best efforts to demonstrate that 'monolingual' and 'bilingual' groups are compatible in all other respects except

language experience, there may be hidden variables that separate the two groups (Diaz, 1983). While an individual differences approach can't help solve this directly, at least it can demonstrate that groupings of participants emerge from the data which align with notions of 'monolingualism' and 'bilingualism'.

This approach is good for exploring hypotheses. However, the individual differences approach does not mean a lack of commitment to an experimental hypothesis. A priori hypotheses are still useful guides, but the focus shifts from predictions about categories of people to the causes of those categories. The traditional approach makes predictions along the lines of 'bilinguals and monolinguals differ in X behaviour'. The individual differences approach makes predictions like 'increased variation in the input affects X behaviour'. Indeed, many studies could describe their outcomes without reference to the coarser and more vague categories of 'monolingual' or 'bilingual' at all.

However, refraining from making assumptions about speakers' linguistic experience is not easy. For instance, while Paradis (2011) uses low-level features to describe the linguistic experience of the experimental participants, the parents of the participants are assigned to coarser categories. Furthermore, thinking about how complex populations of speakers interact can be difficult. The cultural evolution approach can offer solutions by using tools from the study of language evolution.

8.6.1.1 Using tools from language evolution

As discussed above, studies of bilingualism have moved away from categorising individuals into high-level categories towards measuring low-level features of many aspects of the input. Some of these aspects relate to properties of the individuals providing the input, such as parents. However, if the new focus of bilingualism research is to study how specific measures of variation affect specific measures of learning, where to stop measuring can be difficult to judge. For example, a child's executive control is affected by the proportions of languages the mother uses. Therefore, an experiment into executive control will want to measure the mother's variation. However, the mother's variation would be affected by *her* parent's variation, so maybe the grandmother's variation should be measured too, and so on. Worse, considering the input of both parents, every generation the analysis steps back in time doubles the number of people that need to be taken into account. Worse still, children receive input from multiple speakers who are connected in complex social structures and overlapping generations.

Of course, attempting these measurements would be very difficult. Typically, only the immediate family of the child are considered. However, this common-sense approach actually has a formal basis which is often used in models of language evolution, namely a Markov chain assumption (see Gardiner, 2009, ch. 3). This assumes that the properties of an infinite chain of states can be analysed by

considering just a local part of the chain, for instance the previous state. Markov chain assumptions have been used in models of language evolution (e.g. Griffiths and Kalish, 2007). As well as recognising that states are affected by a chain of preceding states, Markov analyses also provide a way of linking the observations of a single generation to the general dynamics of the chain (calculating the stable distribution, as used in Griffiths and Kalish, 2007). For instance, from data on the probability of a child acquiring two languages in different social situations, one can calculate the bias a child has for bilingualism. I apply this technique to data on children acquiring minority languages (from De Houwer, 2007) in appendix B, page 201).

As demonstrated by chapters 5 to 7, computational models can be used to help think about conceptual issues surrounding bilingualism. For example, this thesis argues that a central factor in bilingualism is dynamic social structures. As experimental questions become more complex, computational models can help clarify the predictions made by different theories. This is already used, for instance, in experiments on lexical selection (e.g. Calabria et al., 2011). They can also be used to work out predictions or baseline performances in experiments where multiple individuals interact. This is useful when considering the feasibility of a study. For example, for a pilot experiment that simulated the minimal naming game (see section 5.3 in chapter 5 on page 60) with multiple human participants, I used a computational model to work out whether a manipulation of the strategy of a player controlled by computer would be reflected in the performance of the group (see section D.2 in appendix D). The model results showed that the experiment results should provide an insight into the strategies that the human individuals were using.

The field of bilingualism can also inform computational models of language evolution. For example, the bottom up model allowed social structures to change over time, but the social structure was not affected by the linguistic divisions. The literature on bilingualism and sociolinguistics suggests that the social structure adapts to the linguistic identities of individuals, (e.g. individuals who speak similarly form communities, Berndt, 1959; Yallop, 1969; Chagnon, 1968, or cultural practices such as marrying outside your linguistic community, e.g Jackson, 1983). Quillinan et al. (2010) uses agent-based models to explore the kinds of social structures that emerge when individuals preferentially interact with other individuals who speak in a similar way (see also Gong et al., 2004). The two fields can negotiate so that the field of bilingualism suggests reasonable abstractions for computational models, and computational models suggest reasonable baseline behaviour for the kinds of systems that researchers in the field of bilingualism study.

8.6.2 Solution: Recruiting participants

The traditional approach faces problems in recruiting participants. The participant pool may be restricted due to narrow linguistic criteria and many participants may be needed to ensure statistical control between groups. An individual differences approach can help solve these problems.

An individual differences approach exploits all responses from participants, instead of averaging responses within participants (as in some other statistical approaches, such as ANOVAs). This has two implications. First, given the same experiment, an individual differences approach might need fewer participants to gain the same level of statistical power as a traditional approach. Secondly, this means that it is feasible to include participants with a greater range of linguistic experience. The differences in non-experimental variables between individuals can be controlled using statistics.

An individual differences approach also allows unbalanced groups, or groups where there is a continuous range of linguistic experience. For example, the linguistic experience of some participants may have overlapping factors (e.g. some participants who were late learners of a language also use the language on a regular basis). It also allows controlling for conditions or combinations of factors that are not observed in the experiment (e.g. maybe there are no participants who were exposed from birth *and* did not have two bilingual parents). This is useful if studies of bilingualism want to explore a greater range of bilingual experience.

However, the individual differences approach does not mean that an experiment can use any sample of the population. Participants are needed who exhibit diversity in the relevant factor or factors under investigation, so researchers must still exercise their judgement in selecting participants. In many cases researchers could use the same populations, but analyse them differently. An experiment where groups are used to control for other factors may be more convincing, but an individual differences approach may be useful when exploring new hypotheses or when experiment participants are difficult to find.

8.6.2.1 Cumulative results

If experiments collect a lot of low-level, context-independent data, then experiments may be able build on each other's data more easily. Given the same experimental set-up, it may be valid to include data from previous experiments in the current one. This would mean that there may be no need to recruit additional control participants. Furthermore, statistical inference would become increasingly powerful as results accumulated.

In order to realise this potential, studies would need to collect compatible, low-level, context-independent data. This is already suggested by the cultural evolution approach. Detailed questionnaires about linguistic experience are already in

use (see De Houwer, 2009; Gutiérrez-Clellen and Kreiter, 2003; Jia and Aaronson, 2003; Paradis, 2011; Unsworth, 2011).

An individual differences approach allows freedom from constraints such as only including participants that fit a narrow definition of bilingualism and being able to compare data from different experiments. This could broaden the scope of studies of bilingualism.

8.6.3 Solution: Conclusions in terms of mechanisms

As chapter 3 suggests, the concept of bilingualism involves many different considerations apart from just the knowledge of two languages. As well as features of proficiency and context of learning, it involves the knowledge of who to use those languages with, which often goes along with experience of a wider variety of people, cultures or places. That is, a measure of the number of countries lived in, for example, can be seen as part of the measure of bilingualism. Rather than impose an idea of what bilingualism is onto the research, an individual differences approach can let the relevant features that define bilingualism emerge from the differences in the experimental task.

The conclusions from this kind of approach will be described in terms of lower-level measures. For example, rather than saying that bilinguals have better cognitive control than monolinguals, a conclusion could say that enhanced cognitive control occurs with exposure to contexts where individuals must inhibit certain responses based on social variables. The latter statement better reflects the hypothesised mechanism. This is a reductionist approach, which might be necessary if the concepts behind ‘bilingualism’ are changing or are to be integrated into other fields such as cultural evolution.

8.7 Implications

The solutions above allow experiments to make more valid assumptions about the linguistic background of participants, recruit a wider range of participants and link properties of the linguistic input directly to learning mechanisms. Below I discuss some possible studies that fall out of the cultural evolution approach.

8.7.1 Answering relevant questions

The cultural evolution approach suggests collecting low-level data from a wide range of bilingual experiences. This could help give more concrete answers to questions from outside of academia. Parents who raise bilingual children often express anxiety about whether their child will acquire a given language and the best methods for ensuring bilingual acquisition (e.g. Petitto and Kovelman, 2003;

De Houwer, 2007). The *Bilingualism Matters* Network⁸ aims to provide parents with relevant, accurate information about bilingual language acquisition, based on current research. However, parents' questions often describe a quite specific context, then ask how they can improve the chances of their child becoming fully bilingual. Here's one example posted online to an online *Bilingualism Matters* forum⁹:

"I live in a remote place in Scotland with my Scottish husband and our two kids. Although I would like my kids to be able to speak Dutch (my mother tongue) I find it really difficult to be consistent and persistent as I am their only source of the language. I now wonder if I left it too late for them to really pick up Dutch. They are 5 and 8. My 8-year old's Dutch is ok-ish as he lived in Holland for the first 3 years of his life but my daughter's Dutch is very dormant. The other issue that I find difficult is that Dutch is not a major language in the world. It doesn't have the same importance as for example Spanish or French. Please advice [sic]."

The features here include age of acquisition, exposure to multiple speakers, exposure to a native culture, access to linguistic resources, consistency in parental language use and the perception and prestige of the language. Current experiments are usually set up to answer one or two of these features in a very general way. The majority of advice to parents from the 'Bilingualism Matters' group is to stop worrying, citing general findings that children are flexible and sensitive learners.

The cultural evolution approach could lead to more specific answers. It suggested that experiments should use low-level, context-independent measures to build cumulative databases of compatible experiment results from a wide range of bilingual experiences. These data could be used to create a statistical model of the effects on and results of bilingualism. For example, one could build a model so that parents could enter relevant factors of their situation and be given the likelihood of their child acquiring proficiency in a minority language. The model could also predict other areas of bilingual advantage such as executive control (e.g. Bialystok and Craik, 2010), reading skills (e.g. Edwards and Christophersen, 2011) and interpersonal skills such as empathy (e.g. Dewaele and Wei, 2012). The model would exploit the individual differences of experiment participants to fit predictions as close to the situation of the parents as is possible. Doing this with the traditional approach (presuming results were compatible) would involve first assigning the child in question to a categorical group, which might not be suitable.

More importantly, the model could suggest key areas where a change would have the biggest impact. This might depend on the context. For example, in one situation increasing the amount of exposure to other speakers might be more likely

⁸<http://bilingualism-matters.org.uk/>

⁹<https://www.facebook.com/groups/72220629226/>

to lead to good acquisition while in another context a ‘one-parent-one-language’ approach might be more likely to yield greater benefits.

This kind of model could have a practical impact on bilingual communities in the short-term. It would also be useful in terms of understanding how linguistic variation, learning biases and social structures coevolve. This is a speculative project, not least because it would require collaboration and sharing of data between institutions and researchers. However, sharing data has been productive in other areas of language acquisition such as the CHILDES database (MacWhinney, 2000). There are also several projects that are addressing the theoretical and practical problems involved in building collaborative databases for language acquisition research (e.g. McCue et al., 2007; Steinhart et al., 2008; Lust et al., 2010).

8.7.2 Artificial language learning

A major challenge for many studies of bilingualism is finding participants with the relevant language background, especially when considering infants. However, finding participants with the relevant language background for some language evolution studies is essentially impossible, since they would require humans without knowledge of language. Despite this, there are a number of experimental techniques that have been used to study language evolution. Artificial language learning paradigms have been used to investigate how individuals learn multiple conflicting grammars (e.g. Nation and McLaughlin, 1986; Nayak et al., 1990; McLaughlin and Nayak, 1989). Learners who have experience with a greater number of natural languages learn artificial grammars better and faster (Kemp, 2001) and also use a greater range of learning strategies (Kemp, 2007). Participants can also adapt to relevant sources of information in the input (e.g. Gómez, 2002; Saffran et al., 1999; Smith et al., 2011).

Artificial languages have also been used in the iterated learning paradigm (Kirby et al., 2008; Tamariz et al., 2012; Perfors and Navarro, 2011). Participants learn an artificial language, and then are asked to reproduce it. Their reproduction then becomes the input for the next generation. This happens repeatedly to create a chain of learners. The way the language changes can be tracked over time. If the study of bilingualism can be reformulated as the study of how linguistic signals and social structures coevolve, then this paradigm could be used. Some preliminary investigations of the impact of social structure in this paradigm suggest that it has an impact on the emergence of structure in the linguistic system (Tamariz et al., 2012; Line, 2010; Tan and Fay, 2011).

There are two ways artificial languages may be used in studies of bilingualism. These involve testing participants with profiles of interest, and priming participants to have profiles of interest. The former participant group can be studied using iterated learning. Griffiths et al. (2008) demonstrate that chains of iterated

learning can reveal the inductive biases of individuals (also reviewed in section 7.4.1 in chapter 7). Participants were exposed to a learning task and then tested to produce data for the next participant in the chain. While the change between individuals was slight, over many iterations the distribution of the data converged to reflect the learning biases of participants¹⁰. It would be therefore possible, for example, to estimate in real learners the strength of their cognitive bias for expecting variation in the input. As long as the data that a participant sees is the result of the output of a previous learner, the eventual distribution of the data should be informative for the hypothesised biases. This means that it may be possible to run this kind of experiment with infants. It may be possible to track the strength of particular learning biases of ‘monolinguals’ or ‘bilinguals’ over the course of development to examine how cognitive biases adapt to different types of input.

Alternatively, it may be possible to ‘create’ bilinguals in the lab. That is, expose (monolingual) adults to an artificial language learning environment that will prime a cognitive profile compatible with real bilingual profiles. For example, adults completing a cross-situational word learning task¹¹ adapt their learning strategies according to whether words are presented in blocks or interleaved (Smith et al., 2011). One could imagine, then, constructing an artificial language learning experiment where participants had to learn two languages, L1 and L2 and then were then tested on one (see figure 8.3). The order of learning could be manipulated to prime different profiles: L1 first then L2 second, then tested on L2 would prime an L2 learner profile. L1 first then L2 second, then tested on L1 would prime an attrited speaker. Being exposed to interleaved rounds of L1 and L2, then tested on L1 would prime a balanced bilingual profile. It would be interesting to see if similar effects to the ones mentioned above could be obtained in the short term. In a similar way, it may be possible to prime executive control profiles in the short term, helping to explore the questions raised about how language experience and executive control are related (see above and section 5.9.3).

The advantages of this kind of experiment are, firstly, that it allows very detailed control over the language experience of the participant. Secondly, finding eligible participants is much easier and experiments are quicker to run. Finally, it is possible to construct control conditions that might be very rare or not actually

¹⁰This task was a function learning task where participants were exposed to images of two bars whose lengths covaried. They were asked to learn the function that determined the relationship between the bar lengths. Then they were asked to produce the correct length of one bar given the length of another. Regardless of the function in the initial data, the data converged to a linear (usually positive) relationship, suggesting a bias towards positive linear relationships.

¹¹Cross-situational word learning tasks involve participants learning mappings between novel words and novel objects by being exposed to many instances of a novel word paired with a set of novel objects. There is referential uncertainty in any one of the exposures, but by combining information across exposures, the correct mapping can be deduced.

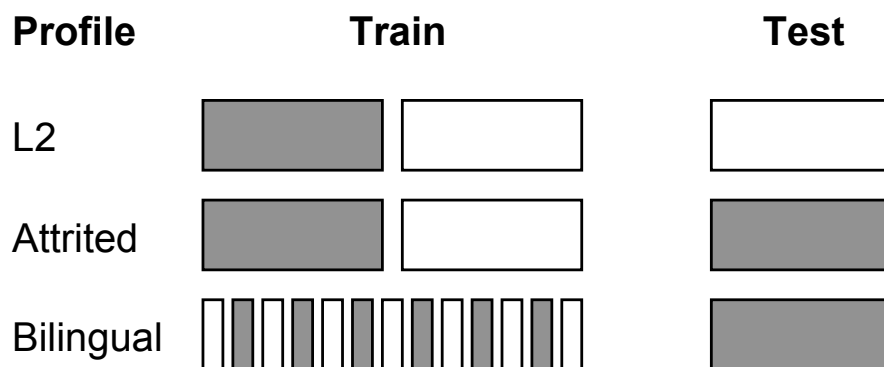


Figure 8.3: Methods to prime different cognitive profiles in the lab, based on training and testing different artificial languages in different orders. Grey and white boxes indicate blocks of training or testing in different artificial languages.

exist in the world¹². Artificial language learning experiments trade off realism for control. This is similar in many ways to the trade off between real experiments and simulations, and results from each can mutually support each other (see Scott-Phillips and Kirby, 2010 and Cornish et al., 2009, p. 199). In the next section I present a pre-pilot experiment which demonstrates that creating bilinguals in the lab may be feasible.

8.7.2.1 Creating bilinguals in the lab: a proof of concept

Bilinguals have been shown to have better executive control than monolinguals (e.g. Bialystok and Craik, 2010; Hernández et al., 2010; Treccani et al., 2009). Children raised bilingually have more linguistic variation they need to inhibit, but might also be exposed to more variation in other non-linguistic cultural aspects. Controlling for this is difficult. However, while many studies use previous experience as an independent variable, it could be used as the dependent variable. It might be possible to prime executive control in the short-term, therefore ‘creating’ participants with the cognitive profiles similar to bilinguals, but without necessarily knowledge of two languages. These could serve as useful controls. An informal experiment tested whether this was possible.

Method: 18 participants were recruited¹³ and split evenly into two groups: A switching group and a control group. The participants in the switching group played alternating rounds of ‘rock-paper-scissors’ (see “Rock-paper-scissors”, n.d.)

¹²For instance: where utterances are not uttered by specific individuals (e.g. Mackey, 1965); a system where responses must be in a different language to the requesting language (e.g. Lo, 2002); where communicated strings are covertly scrambled (Hurford, 1999); where objects have impossible properties (Scott-Phillips et al., 2010)

¹³The participants were from diverse language backgrounds. The purpose of this informal experiment was to demonstrate that the concept worked, so no detail about the linguistic background of individuals was taken. Obviously, given the arguments of this chapter, a full experiment would control for these factors. While this experiment is informal, it provides more support for the ‘creating bilinguals in the lab’ proposal than mere speculation.

and ‘odds or evens’ (see “Odds or Evens”, n.d.) in pairs for five minutes while the non-switching group played only ‘rock-paper-scissors’¹⁴. The rules of the game were explained to the participants before starting. Tokens were used to help participants keep score. The participants then completed an inhibition task to measure their executive control performance. The inhibition task involved writing a list of nouns (in their native language) where no adjacent pairs had an obvious semantic connection. Participants wrote as many as they could think of in a minute. It was predicted that the switching group will do better at the inhibition task.

Results: The switching group did significantly better at the inhibition task than the control group (mean for switching group = 17.3 items, mean for control group = 12.7 items, $t = -2.7$, $p = 0.01$, see figure 8.4). The qualitative results were not altered by taking into account the participant’s score for the priming games. Priming a non-linguistic executive control task affected linguistic inhibition in the short-term. If this effect can be shown to be robust, then it might suggest that the bilingual advantage in cognitive control stems from experiences in non-linguistic domains. Bilingual children may be exposed to a greater range of non-linguistic cultural stimuli (e.g. because they may have parents or extended family from different cultural backgrounds). If this is the case, then the results suggest that non-linguistic executive control advantages might stem from variation in non-linguistic domains, thus resolving part of Sorace (2011a)’s ‘bilingual paradox’ (bilingual children appear to exhibit benefits in non-linguistic domains, but deficits in some areas of language competence).

8.7.3 Comparative approach

If the central question about bilingualism is how input factors affect learning mechanisms which affect communicative output, then the scope of relevant objects of study could extend beyond just humans. That is, we might be able to use insights from studies of how non-human animals learn and communicate to inform theories of bilingualism in humans. This comparative approach (Hauser et al., 2002) recognises an evolutionary relationship between the cognitive mechanisms of humans and other animals. There is already some work along these lines. For example, Hayashi and Matsuzawa (2003) use the same experimental conditions for human and chimpanzee infants (including the same experimenters, stimuli and laboratory), observing differences in the way the two species play and interact with their parents.

The central requirement for a bilingual ability, according to this thesis, is the ability to condition linguistic variation on social variables. As chapter 2 suggested,

¹⁴These are popular children’s games based on co-ordination and chance. For example, in “odds or evens”, one player takes the role of ‘odds’ and one the role of ‘evens’. Both players choose to display either one or two fingers at the same time. If the sum of the fingers is odd, the ‘odds’ player wins. If the sum of the fingers is even, the ‘evens’ player wins.

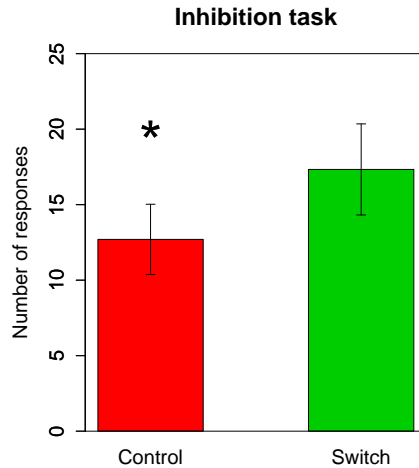


Figure 8.4: Mean number of responses to the inhibition task by group with 95% confidence intervals. The group who switched between two tasks prior to the inhibition task gave significantly more responses on average.

other species *primarily* learn how signals are conditioned on social variables, and it is the conditioning on other semantic variables (identity of tertiary objects, objects displaced in space or time) that is difficult. This might suggest that the basis of a ‘bilingual’ ability is evolutionary very old, and amenable to a comparative approach.

One area that has been studied in both bilingualism research and animal behaviour is word learning and mutual exclusivity. Human children can use the principle of mutual exclusivity to disambiguate the referent of an utterance (Markman and Wachtel, 1988; Markman, 1992). However, children raised bilingually do not consistently use the mutual exclusivity principle (Au and Glusman, 1990; Merriman and Kutlesic, 1993; Healey and Skarabela, 2009; Byers-Heinlein and Werker, 2009a; Houston-Price et al., 2010). Recent research in animal behaviour has shown that non-human animals are capable of using mutual exclusivity, too, including dogs (Kaminski et al., 2004; Pilley and Reid, 2011¹⁵) and parrots (Pepperberg and Wilcox, 2000). For a direct comparison of animals and children, see Grassmann et al. (in press) and Markman and Abelev (2004) discusses how results from experiments with dogs informs theories of mutual exclusivity in humans. These results might suggest that the use of the mutual exclusivity principle is not the ‘default’, but just a particular heuristic that is learned in the right situation.

¹⁵Alliston Reid suggests that Chaser could ‘learn by exclusion’ (eliminate objects with known labels as the target of a novel label), but also learned multiple words for the same object (a proper-noun like ‘bunny’, but also a common noun like ‘toy’), which goes against mutual exclusivity (see response to my blog post <http://replicatedtypo.com/dog-exhibits-mutual-exclusivity-bias>).

Other experiments could also have an impact on theories of mutual exclusivity. For example, Inoue and Matsuzawa (2007) found that chimpanzees can correctly remember the location of 9 randomly arranged numerals displayed for 210ms - shorter than an average human eye saccade. Humans, however, perform poorly at this task. Matsuzawa (2012) suggests a semantic link hypothesis: while chimps have good visual, eidetic memory, humans are good at symbolic associations (see Roberts and Quillinan, in press, for an empirical test of this hypothesis). The extra information in the semantic, linguistic links that humans possess increase the load on working memory and make this task difficult for them. Indeed, chimpanzees trained to use a basic language also performed poorly. This could be linked with the effect found in Mather et al. (2010), which shows that 9-21 month old human infants take longer to develop a visual novelty preference for objects when they are presented simultaneously with auditory labels. If the effect is additive, so that multiple labels delay the novelty preference further, then this might explain the difference in mutual exclusivity behaviour. Rather than a pragmatic response, bilingual children's tendency to assign novel labels to novel and known objects equally often may be a delayed novelty preference, underpinned by a processing explanation. This insight could affect the kinds of controls that are considered in research on word learning.

8.8 Conclusion

Traditional approaches assume a qualitative, categorical difference between monolinguals and bilinguals. This thesis has argued that there are conceptual flaws with this approach which lead to some problems with research into bilingualism. This chapter presented a cultural evolution approach to bilingualism, which suggests exploiting the variation in individual differences to broaden the scope of bilingualism studies. Indeed, some recent research takes an individual differences approach using low-level measurements and avoiding placing participants into discrete categories.

The question that arises out of this approach is how input factors affect the way individuals learn and how this, in turn, affects features of the linguistic output. The object of study, then, is not individuals with a particular type of linguistic experience, but the way a wide range of linguistic experience and cultural experience interact. This brings the research questions of the field of bilingualism closer to those of the field of language evolution, which also considers how linguistic variation, learning biases and social structures coevolve.

The sections above suggest that the scope of the field of bilingualism could be increased by taking a cultural evolution approach which promotes the integration of detailed linguistic measures with new experimental paradigms, computational models and comparative evidence. This will necessarily mean interdisciplinary

work, and the need for collaboration across fields. While the bulk of this thesis has shown how bilingualism is relevant for language evolution, it should also now be obvious that language evolution is relevant for studies of bilingualism.

CHAPTER 9

CONCLUSION

“Behind these new ideas about how society should be managed was a model of the individual as a rational, calculating machine whose self-interested behaviour could be analysed by numbers. They had made an assumption that we were like that, simply in order to make their equations and their models work. But what was now rising up was a powerful scientific proof that this was not just an assumption. That we really were computing machines, guided by numbers.”

Curtis (2007)

9.1 An evolutionary approach to Bilingualism

This thesis set out to explore bilingualism from an evolutionary perspective. It quickly became clear that traditional concepts used to define bilingualism become problematic when thinking about biological and cultural change in the long term. Through a consideration of the abilities of other species, and through modelling cultural evolution, a concrete definition of bilingualism emerged which represents this thesis’s contribution to knowledge. This definition of bilingualism is the amount of linguistic optionality that is conditioned of social factors. The ability to condition optionality on social factors is argued to be evolutionarily old. This implies that, instead of seeing bilingualism as a peripheral ability to be studied after monolingualism is well understood, bilingualism can be a central part of the story of language evolution.

The thesis presents two types of model that look at the dynamics of bilingual cultural evolution. Top down models assume that languages are discrete, monolithic, stable entities and individuals have a specific, rational learning mechanism for learning multiple languages. These models suggest that the key factors that affect the dynamics of the prevalence of bilingualism in a population are the cognitive bias towards learning multiple languages in each individual and the structure of the transmission chain. Although allowing computational agents to learn more than one language might seem like a trivial task, it requires qualitative changes to the models. A bottom up model, on the other hand, assumes that languages dynamically emerge from use and that individuals have general learning mechanisms that condition linguistic variation on salient semantic variables. The key factor that affect bilingualism in this model is dynamic changes to the social structure.

Previous chapters suggested that there were three resolutions to the ‘bilingual paradox’: If learning two languages is harder than learning one, why is bilingualism so prevalent? First, an evolutionary approach to bilingualism (chapter 2) suggests that the paradox is using the wrong unit of analysis. Rather than the number of ‘languages’, it is the complexity and saliency of the conditioning factors that poses a challenge to learning.

The other two resolutions come from different types of computational model. The top down model suggests that the inference in the paradox does not hold. Population-level phenomena (e.g. the number of bilingual speakers) are not straightforwardly linked to individual learning biases (e.g. an expectation there will only be one language in the input). Cultural transmission can lead to distributions at the population level that exaggerate the individual bias, or in some cases, go completely against them.

In contrast, the bottom up model suggests that the paradox considers both the wrong unit of analysis and the wrong conception of bilingualism. That is, bilingualism is not just a measure of the variation in linguistic variables in a population, but how those variables are conditioned on social variables. If social variables are salient conditioning factors, then bilingualism emerges. This fits with the evidence from the literature on bilingualism that shows that children are adept at learning multiple languages simultaneously due to flexible learning mechanisms. However, a key factor that leads to salient social conditioning factors is dynamic social structures. Social distinctions in stable populations may be hard to maintain, and the amount of bilingualism will decline, not due to competing variants, but due to a lack of variation at the social level.

The results of the bottom up model suggests a change to research questions in language evolution and language acquisition. Rather than asking why there is so much linguistic variation, evolutionary linguists could equally ask why the language acquisition capacity is so flexible. Similarly, researchers in the field of bilingualism could shift from asking ‘what are the differences between monolinguals and bilinguals?’ to ‘How do linguistic variation, learning biases and social structures coevolve?’

The differences between the results of the two types of model demonstrate that the tools that are suitable for studying one kind of problem (e.g. cultural evolution of linguistic variants) are not necessarily suitable for studying another (e.g. the cultural evolution of bilingualism). The top down models are not fundamentally opposed to the bottom up approach, but the constraints of the top down method tend to bias researchers towards making certain kinds of simplifying assumptions. These will be specific to the research question, so applying them to other questions may not always be valid. This suggests that researchers must continually uncover and test the assumptions that they make. It may be difficult to predict the wider implications of theories about language evolution and

bilingualism.

9.2 Wider implications

This thesis suggests that applying assumptions designed for one research question to a different research question may not be valid. Applying assumptions across fields may also be problematic, so researchers should be careful to articulate their assumptions behind and limits of their models. For example, in the 1950s, John Nash developed mathematical models of strategising called Game Theory which assumed that individuals were self-interested, rational beings (e.g. Nash, 1950). The legacy of this theory ran deep, in theories of biology (Smith and Price, 1973) and theories of economics (e.g. Shapiro, 1989) but also in political strategies (Downs, 1957), social policy (e.g. Kahan et al., 1992), foreign policy (see Mintz and DeRouen Jr, 2010, ch. 4), morality (e.g. Sober and Wilson, 1999) and even psychiatry (e.g. Colman and Wilson, 2011). In a series of television documentaries, Curtis (2007, 2011) argued that these have had damaging consequences on society. Recently, Nash has stated that he “may have over-emphasised rationality ... on the part of humans” (Curtis, 2007).

Many models used to explore cultural evolution assume that humans are rational. These are attempts to create a science of how culture changes and evolves. This thesis has argued that assumptions about rationality and about the ideal speaker and community can affect the direction of research, leading to a disregard for the reality of the individual and a view of bilingualism as a paradoxical, minor, even unwanted situation.

This thesis has demonstrated that bilingualism is a complex concept which often relies on identity, history and politics. In fact, the linguistic uniformity between individuals is independent from the amount of linguistic variation within individuals. Everyone can agree that there’s more than one way to say something. A model with a non-rational, flexible learning mechanism demonstrates that, while populations typically strive towards uniformity, social factors shape what features become important parts of that uniformity. It is important that researchers’ assumptions are constantly uncovered, clearly defined, explored and re-evaluated. There’s no telling where these assumptions may be reapplied.

9.2.1 Language Policy

Perhaps one obvious area of possible impact is language policy. Language policies can have a big impact on societies and are heavily politicised. For instance, Pifer (1979) claims of bilingual education policies that “few other educational experiments in recent years have managed to arouse such passionate debate” (Pifer, 1979, p. 4). Researchers in the field of language policy have been drawing on work from language evolution. Indeed, in the introduction to *Language Policy*,

Spolsky discusses language evolution, citing some of the literature reviewed in this thesis and concluding that dynamic social structures are an important factor in language policy:

“Language evolution is to be explained not just by small random variation strengthened by geographical isolation, but also by including functional and social selection. Nettle (1999:79) proposes that different “ecological regimes favored different kinds of social networks, which in turn produce different-sized linguistic groups.” The activating factor in his model is “ecological risk,” managed by non-industrial societies by forming social networks that reduce diversity as people communicate with each other. The greater the ecological risk, the more interaction and so the fewer languages there will be in a country of a given size and population. The change from hunter-gatherers to farmers and herders reduced linguistic diversity, as did European expansion and industrialization. It is social policy rather than language policy that is needed to maintain it.”

(Spolsky, 2004, p. 7-8)

In another example, Grundy (2007) relates the iterated learning model to methods of teaching English as a foreign language. Grundy “considers the significance of Kirby’s iterated learning account of language evolution”, suggesting that there are “striking parallels between the evolution of pragmatic inference and choices in second language teaching methodology” (Grundy, 2007, p. 219-220) and concludes that the interaction between learning biases and cultural evolution would naturally favour a “use-in-order-to-learn” approach to teaching English as a foreign language. Grundy also argues that “the current status of English [as a global lingua franca] is not only a consequence of political and social development but also of language evolution and iterated learning” (Grundy, 2007, p. 219).

Furthermore, rational approaches to linguistics have been explicitly linked with rational approaches in language policy (see Ingram, 2003, p. 11). Some extend this rational approach to language learning to implications about the dynamics of bilingual cultural transmission. For example, Mackey suggests that “a self-sufficient bilingual community has no reason to remain bilingual, since a closed community in which everyone is fluent in two languages could get along just as well with one language” (Mackey, 2000, p. 22). The bottom up model presented in this thesis demonstrated that bilingualism decreases when social structures are stable. However, the result of the model does not endorse Mackey’s statement. The model was extremely simple and contained no aspects of cultural identity, prestige, economics, politics and so on. It was a ‘null’ model that demonstrated some basic factors of bilingual cultural evolution, and used a concept of bilingualism that is not immediately analogous with situations recognised as ‘bilingual’ in the real world. The results of the model should not be taken as a proof or validation of policies that limit linguistic variation. In fact, they suggest quite the opposite, as will now be outlined.

The aims of a language policy should be to maximise people's access to education and to maximise cohesion in order to increase co-operation and national development (Ingram, 2003; Kaplan, 2008). Some models discussed in this dissertation show how populations achieve cohesion. However, many have an implicit assumption that variants are in competition and increasing cohesion involves decreasing variation (see section 5.3 in chapter 5). Some models see the choice to learn languages as rational and self-interested, based on personal gain (e.g. Iriberry and Uriarte, 2012). Those belonging to the majority group therefore don't have a rational reason to learn the languages of minority groups. The language policy that makes sense given these assumptions is to emphasise the incentives the minority groups have to learn the majority language.

However, there is another possible route to cohesion which involves maintaining the levels of variation, but altering the distribution of variation so that everyone speaks all languages. A policy that makes sense under this approach is to give everybody an equal opportunity to learn all the languages, and to emphasise the incentives the majority group has to learn the minority language. This might seem as an unrealistic policy, but is in fact being used to promote languages like Welsh, Scottish Gaelic and multilingual schools in the UK and Europe. For instance, the Bilingualism Matters (<http://www.bilingualism-matters.org.uk/>) organisation encourages parents to raise their children bilingually for the benefits to cognition (e.g. Bialystok and Craik, 2010; Bialystok, 2011; Lauchlan et al., 2012), among others. Indeed, while some assume that linguistic diversity in education can only be disruptive (e.g. Shipton, 2008; Penman, 2010), a UK school with pupils from 72 language backgrounds was seeing 92% of its students receive 5 A-levels at A*-C level recently (Sollis, 2010). While some areas have less linguistic diversity to take advantage of, similar benefits may be achieved by educating people about linguistic diversity through the study of linguistics in schools (see Arnold, 2012). Multilingual language policies have been shown to have advantages over monolingual policies, including advantages beyond the ability to communicate (e.g. Hornberger, 2002, 2009).

Some models make explicit links to language policy. Abrams and Strogatz suggest that their model "may be useful in the design and evaluation of language-preservation programs" (Abrams and Strogatz, 2003, p. 900, see section 5.2 in chapter 5). However, this statement has been criticised (Fernando et al., 2010), not least because bilingualism was not possible in their model (see section 5.2), and their only suggestion - to raise the prestige of the minority language - is undermined by the fact that simpler models without prestige can produce the same results (Reali and Griffiths, 2010, see section 5.9.1 in chapter 5).

In another example, Iriberry and Uriarte (2012) use a game theory approach (a Hawk-Dove game) to the stability of bilingualism. They conclude that reducing the uncertainty about a bilingual's identity can increase stability of minority

language, and expressly link this to possible language policies. However, this may not be a realistic solution, as they also note that “a language policy consisting of marking or labelling people to denote their bilingual nature could be an additional source of conflict” (Iriberry and Uriarte, 2012, p. 21).

Fernando et al. (2010) demonstrate the difference between extending an interpretation of a model to implications for language policy and building models whose purpose is to explore language policies. Fernando et al. present a mathematical model of the transmission dynamics of a community with a high-status language and a low-status language. The baseline behaviour of the model is that low-status language eventually becomes extinct. This is compared with the dynamics under 3 types of intervention: (1) promoting the status of the low-status language so that learners are more motivated to learn it; (2) using government programs to increase exposure to the low-status language and (3) formal teaching of the low-status language to children who speak the high-status language. These interventions can maintain the low-status language in the population. The most effective strategy is to increase the amount of exposure to the low-status language (2), followed by teaching the low-status language to children who speak the high-status language (3). Increasing the status of the low-status language (1) had little effect unless the change was drastic, in contrast with the conclusions of other models (Abrams and Strogatz, 2003). The solution involves increasing the number of bilingual speakers, rather than maintaining a monolingual low-status language community. Fernando et al. actually suggest that governments could estimate their parameter settings and use the model to decide which intervention strategy to invest in.

In summary, the models discussed in this thesis are very abstract, and it is easy to re-interpret them as representing cultural factors other than language. Researchers should be explicit about the assumptions they are making and the scope of their models since it is difficult to predict where they will be re-applied.

9.3 Future work

This thesis presented two types of model with different assumptions and demonstrated that they can lead to different conclusions. The most productive model given the research questions of this thesis was the bottom up model. As mentioned in chapter 7, there are still many aspects of this model that could be explored, such as manipulating the environment, modelling migration between groups, or allowing the cultural identity of individuals to be influenced by their behaviour. As suggested in appendix A.1, there are options for extending the model so, for instance, the linguistic signal has more than one dimension. Manipulating the learning biases of individuals was also not explored to a great extent (although see manipulations of the stepwise information criterion, appendix C).

Although the bottom up model suited the current research questions, there is no reason why top down models could not model the same process. I have argued that the constraints of top down models may lead researchers towards making certain simplifying assumptions initially, but as chapter 5 demonstrated, top down models are becoming increasingly sophisticated. Since top down techniques might yield more robust results, it is hoped that the bottom up model in this thesis can highlight the kinds of considerations that needed to make a valid top down model of bilingualism.

The conception of bilingualism used in this thesis is intended to be one that can unite different fields of linguistics under a common research question. I have argued that language evolution and bilingualism have much in common. I have also suggested that theories of sociolinguistics such as audience design have a very compatible view of language systems, but it would be interesting to work out further parallels between language evolution and sociolinguistics. Since cultural identity involves geography, history, politics and the psychology of identity, it is possible that many fields could contribute towards the question of how linguistic (or cultural) variation, learning biases and social structures coevolve.

9.4 Conclusion

Bilingualism is best thought about as the amount of linguistic optionality that is conditioned on social variables. Rather than an atypical experience or an ancillary object of study, bilingualism can be seen as a central feature of language use and a central part the story of language evolution. This opens up the scope of the object of study for both the fields of bilingualism and language evolution. This thesis suggests that researchers in both fields are converging on similar questions and may have methods, tools and data to share. Researchers should keep an open mind about the ability to learn multiple languages, but also multiple approaches to studying them.

Appendices

APPENDIX A

LINEAR REGRESSION

A.1 Linear regression

Linear regression attempts to explain the variance in a dataset by assuming the dynamics are linear. For each independent variable $x_1, x_2 \dots x_n$ there is a slope β and an error or intercept ε which is fitted to a variable y which is to be explained.

$$y_i = \beta_1 x_1 + \varepsilon_1 + \beta_2 x_2 + \varepsilon_2 \dots + \beta_n x_n + \varepsilon_n \quad (\text{A.1})$$

Commonly, the ε terms are collected under a single error term. The model presented in chapter 7 would like to capture the ability of language learners to detect which semantic variables are important in explaining the variation in a speech signal. That is, not all variables are necessarily entered into the model. One may represent this by setting β for variables that are not to be entered to zero. However, this is slightly misleading, since values of β near to zero are not necessarily unimportant. A variable with a small slope might explain a lot of the variance while a variable with a steep slope might explain little. More commonly, variables are removed altogether from the model. This means that, as well as each variable having an optimal slope and intercept, there is a space of models containing all possible combinations of variables.

Stepwise regression is a method of searching this space for the best model (a set of variables which have been identified as important conditioning factors for the dependent variable). There are many criteria for comparing models. One consideration is the size of the model. Models with fewer variables may be simpler (low variance), but they may also miss some important explanatory variables (high bias). Models with many variables may cause overfitting. In this model, this may be analogous to the tension between a language that is learnable (simple but also expressive (accounts for many aspects of meaning)). Another consideration in model selection is the amount of variation explained. Adding more variables to a model will improve its accuracy.

However, we would like to take into account both the accuracy and simplicity of the model. We don't want to add a variable to the model unless it improves the results significantly. Information criteria is used to compare models in this way. For a set of nested models (where the variables used in each successive model is a sub-set of the previous model), information criteria can be used to measure the improvement of the model after each variable is added variables. A typical information criterion used in stepwise regression is the Akaike information

criterion (AIC). AIC is calculated as

$$AIC = 2k - 2\ln(L) \tag{A.2}$$

Where k is the number of variables in the model and L is the maximised value of the likelihood function of the model (the likelihood of the data occurring given the model). The model with the lowest AIC is considered the best. This means that AIC rewards accuracy (large L) but punishes models with many variables (large k).

There is a Bayesian information criterion (BIC, also called the Schwarz information criterion or SIC) where the model score is calculated as

$$BIC = \ln(n)k - 2\ln(L) \tag{A.3}$$

where n is the number of data points. This punishes the size of the model more harshly than AIC. Shtatland and Barton (1998) suggest parametrising the two measures as:

$$IC(c) = ck - 2\ln(L) \tag{A.4}$$

Where $c = 0$ is the classical likelihood statistic, $c = 1$ is equivalent to the generalised linear interactive modelling (GLIM) goodness-of-fit, $c = 2$ is AIC and $c = \ln(n)$ is equivalent to BIC. Shtatland and Barton (1998) also point out that it may be better to take the information criteria scores as probabilistic, and focus on a probability distribution over models rather than selecting a single ‘best’ model. They suggest that the models in the ‘window’ between the AIC and BIC selections may provide a sub-set of relevant models for tractability. That is, AIC and BIC may choose different models in the nested space of models and we can consider all the models between these two.

The c variable here could be used as a kind of bias in the model for learnability versus expressivity. With $c = 2$, the more complex models are favoured while higher values of c favour more ‘expressive’ models.

A.1.1 Bayesian linear regression

It is possible to take a rational Bayesian approach to regression (see Griffiths et al., 2009; Kalish et al., 2007, also textbooks on Bayesian statistics e.g. Gelman et al., 1995). Given a space of models, the likelihood of the model is simply how well it predicts the data - this is no different from the likelihood evaluation above (with $c = 0$). In addition to this, a prior probability distribution over the model space is defined, giving a higher *a priori* probability to certain models. This is similar to the term in the information criteria method for punishing large models. However, the Bayesian prior is more flexible and can be used to punish any type of model. The prior is defined as a Gaussian distribution over the β_i and ε_i variables. Learners may have a bias towards a particular variable or a more

general bias, for instance favouring models with steep slopes and low intercepts.

Being able to control a bias over low-level statistical properties of the learner is a useful property. However, it is unclear why one would assume anything else other than a bias for steep slopes and low intercepts. As mentioned above, Shtatland and Barton (1998) suggest that a fully Bayesian approach may not be necessary, especially since the specific priors are initially unknown. A Bayesian approach will not be pursued for now.

A.2 Cluster analysis

Cluster analysis attempts to define categorical clusters from a set of data. For instance, given a set of unlabelled data on vowel quality (F1 and F2), a cluster analysis would reveal the statistically significant clusters. That is, it could ‘learn’ the relevant contrasts in the language.

Cluster analysis has been applied to many problems in language learning (see Vallabha et al., 2007). Vallabha et al. (2007) apply cluster analysis to a real corpus of child-directed speech. They use an extension of the Expectation-Maximisation algorithm¹ that assumes that vowel clusters are defined by a gaussian distribution which themselves have a probability of generating an instance. The model predicts both the number of vowels, the parameters of those vowels and the probabilities of those vowels occurring. In order to generate training and test data, the parameters of the real data were estimated and then used to generate more data. Thus, this paper defines both learning and production.

It would be a relatively simple process to iterate this model. However, the algorithm is more complicated than is required for an abstract model. More straightforward off-the-shelf models are available such as parametrised Gaussian mixture model determined by Estimation-Maximisation. The cluster analysis algorithm searches a space of models defined by a number of categories. Each category is defined by a mean and a variance for each semantic variable. The best model is selected based on information criteria (as discussed in section A.1). This is similar to stepwise linear regression in terms of the kinds of analysis that would be done. For instance, the bilingual measurements discussed in section 7.3.2 would also work for a cluster analysis model. The flexibility of social structure and semantic variables would also be unaffected. However, while the linear regression model will ignore some variables (by dropping them from the model), the cluster analysis model is more continuous, making it less ‘clean’. An iterated cluster analysis model would have a lot in common with Perfors and Navarro (2011)’s experiment discussed in section 7.4.2.

¹The Online Mixture Estimation algorithm developed in Vallabha et al. (2007) is an online Expectation-Maximisation algorithm that calculates its covariance matrix using gradient descent.

The implementation is fairly simple. In fact, here is the code for implementing an iterated cluster analysis model in R (for some results, see figure A.1):

```
library(mclust)
# generate initial data
n = 30
num.params = 3
num.cultural.params = 2
data = data.frame(array(dim=c(n,num.params)))
for(px in 1:num.params){ data[,px] = runif(n)}
# run iteration
for(g in 1:10){
# get best model
dataModel <- mclustModel(data, mclustBIC(data))
# produce new data
dataSim <- sim(modelName = dataModel$modelName,
               parameters = dataModel$parameters, n = nrow(data))
# transmit cultural parameters
data[,1:num.cultural.params] = dataSim[, (1:num.cultural.params)+1]
plot(dataSim[,2:3], col=dataSim[,1])
}
```

The learner finds clusters in the data and reproduces new data based on their model. In this model, there is no distinction between the ‘linguistic’ variables and the semantic variables except that the ‘linguistic’ variables are culturally transmitted. It is also possible to easily define more than one ‘linguistic’ variable, while in the linear regression model having more than one independent variable complicates the model. This model could also be extended to include hierarchical clustering to create hierarchies of clusters.

A.2.1 Learning categories and symbolic links at the same time

Part of the problem with using linear regression is that the links between the linguistic signal and the semantic variables are not arbitrary nor symbolic. One way to address this is to combine the statistical learning properties of cluster analysis with a symbolic learning approach². For instance, the learner could use cluster analysis to determine a set of categorical clusters in the linguistic variables (analogous to words or phonemes) and another set for the symbolic variables (analogous to concepts). It could then construct an exemplar corpus of categorical data given the raw data. For example, a particular case might have linguistic variables that were categorised under word X and semantic variables that were categorised as concept Y. This would mean that word X refers to word Y and the learner could store this example in its memory (see figure A.2). While determining the categories models statistical learning, building the corpus is more like an

²This idea was developed with Márton Sóskuthy

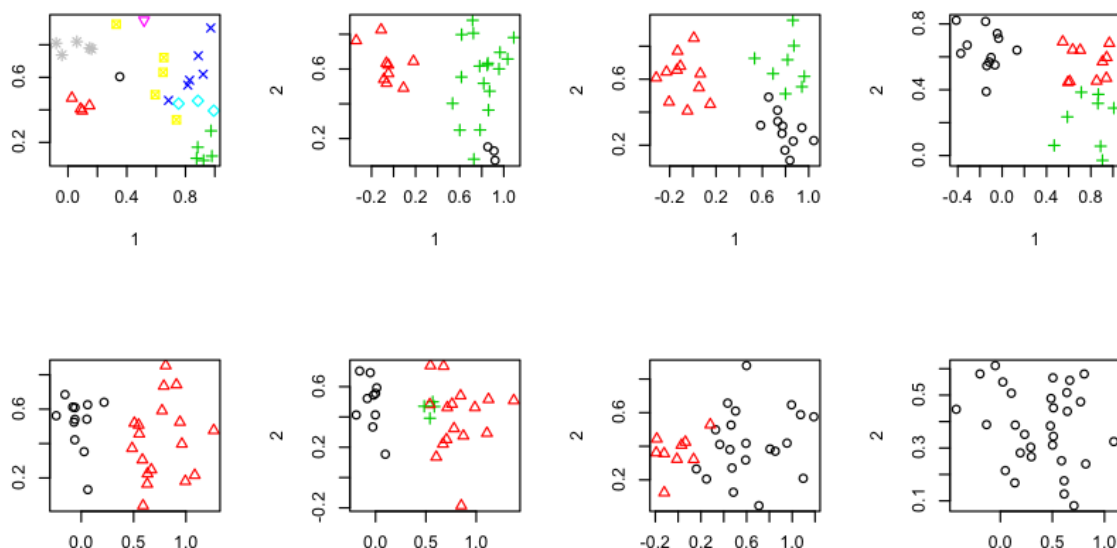


Figure A.1: Categories changing over generations (left to right, top to bottom) in an iterated cluster analysis model. An initial language with many categories reduce to three then two then one category. 2 linguistic variables, 1 non-linguistic variable, $n=30$, initialised with uniformly distributed random variables.

exemplar theory model. The advantage over a simple cluster model is that it has non-arbitrary links between linguistic and semantic variables while the advantage over exemplar models is that categories are estimated rather than pre-specified. Words could also refer to more than one object, bringing this model closer to being able to answer questions about mutual exclusivity.

While the description above suggested defining clusters based on all the semantic data at the same time (multivariate clusters), clusters could also be defined for each variable separately. For instance, a semantic variable that represented shape could divide into two clusters (square versus circle) while the variable that represented colour could divide into three (blue, red, green). The same could be done for the linguistic variables (e.g. vowel height variable and voice onset time variable). The corpus is now much richer with multiple linguistic and semantic variables, allowing the possibility that systematicity could occur (one linguistic category representing a single semantic category).

Obviously, the cluster-exemplar model is more complicated than a linear regression model. However, it still fulfils the requirements outlined at the start of this chapter. Furthermore, it may be more cognitively plausible since the learner has to extract categories out of its input data rather than build a linear model of how a linguistic signal varies with a semantic signal. The move to a discrete exemplar model where the structure of the semantic space influences the language may also be more extendable.

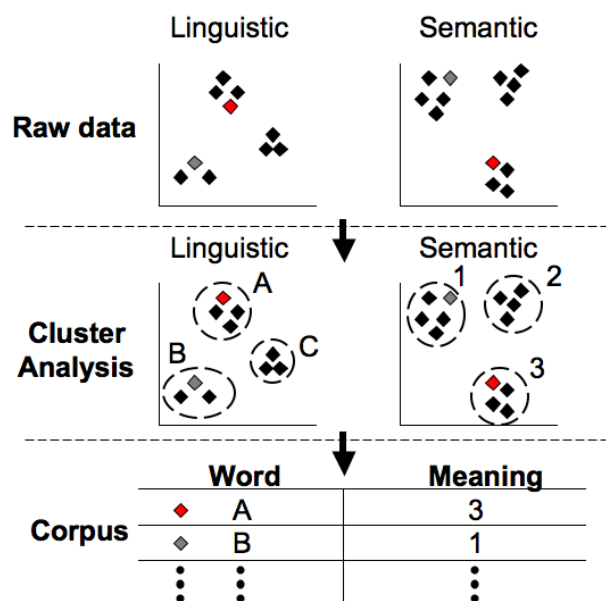


Figure A.2: Learning process in an cluster exemplar model. The grey dots are points generated by the same event (as are the red dots).

A.3 Other modelling techniques

A.3.1 Topographic learning

Topographic mappings are mappings between an input and an output where the distances between input-output pairs are maximised. Ellison (2012) presents a model of cultural evolution where agents learn mappings between signals and lexical meanings using topographic learning. Communities of such agents with overlapping generational turnover rapidly converged on categorical, arbitrary mappings between signals and meanings (Ellison, 2012). The topographic learning mechanism is a potential candidate for replacing the linear regression learning mechanism in the bottom up model in chapter 7, since they are relatively simple and neurologically plausible (Silver and Kastner, 2009).

A.3.2 Iterative regression

Iterated regression is not to be confused with an iterative approach to regression (e.g. Holt and Benfer, 2000) where missing values are filled in by using a regression on other predictors which can then be updated in the same manner.

A.3.3 Linear regularisation

Various methods for introducing biases into the model are possible using linear regularisation methods such as ridge regression (or Tikhonov regularisation) or the LASSO method. There are a lot of similarities with the Bayesian regression model or information criteria approaches, and the choice between them often

depends on application. Since this is a theoretical model, these complications will be put aside for now.

A.3.4 Hidden Markov Models

Language evolution has already been modelled using hidden Markov techniques (e.g. Griffiths and Kalish, 2007) and the current model can be also captured by a hidden Markov model: There are a set of hidden variables (the slope and error parameters of the regression) which must be estimated using only the observed linguistic and semantic variables of the last generation. The model space is therefore traversed by a chain of learners in a Markov chain.

There are reasons for avoiding a Markov analysis for the purposes of the current model, however. The benefit of hidden Markov models is that the dynamics are analysable given certain assumptions. For instance, the Markov chain represents the changing model trajectory of a single individual. However, this model would like to be able to model multiple learners learning from multiple teachers in a changing social structure. In this case it becomes unclear what the current state represents - an individual's language or the community's language or something in between. Furthermore, it's unclear how to model individuals that take different trajectories in the state space. Also, the model would like to be flexible with regards to the observable variables available at each generation, but a hidden Markov model assumes that all states are specified in advance. Therefore, using the hidden Markov model to capture this could be very complicated and perhaps intractable.

Particle filtering is a type of online method for estimating latent variables in a Markov chain where the latent variables are continuous and the state space is not sufficiently well defined. This may help in giving more flexibility with regards to the semantic variables, but still has the problem of representing more than one individual realistically. However, the more advanced methods of estimation may be more useful for engineering solutions in the real world than abstract models of learning and transmission.

A.3.5 Recursive filters

A Kalman filter is a simple implementation of an online recursive Bayesian estimator for a multivariate normal distribution. That is, given a set of noisy independent variables the Kalman filter estimates the dependent variable and updates its estimation as new data is available. It does this by predicting a value for the dependent variable given its current model and current data. It then updates its model when new data appears to a weighted mean of its prediction and the new estimate. The process is therefore online because it only requires the current data and the previous model. Using a particular weighting, this method is guaranteed to produce estimates that are closer to the true values. The method

is similar to a hidden Markov model, except the latent variables are continuous and have a Gaussian distribution.

The Kalman filter has some similarities with the current model. A model is repeatedly updated given only the data available in the last timestep. This is somewhat analogous to new generations receiving input from the previous generation. However, in the Kalman filter model the next generation would also inherit the parameters for the model directly from the last generation. While the Kalman filter may provide some interesting possibilities and has some nice formal properties such as convergence, it won't be pursued here.

APPENDIX B

SIMULATING PARENTAL INPUT PATTERNS

B.1 Introduction

This section applies real data to the Bayesian bilingualism model presented in chapter 6.

De Houwer (2007) conducted a survey of parental input patterns in Flanders leading to the successful acquisition of a minority language. It should be kept in mind that it's likely that many social factors had an impact on the acquisition of the minority language. The two most successful patterns for the transfer of the minority language were where both parents spoke only the minority language, or where one spoke only the minority language and one spoke both minority and majority languages. The pattern least successful for the transfer of the minority language was the opposite of the latter case: where one parent spoke only the majority language and one spoke the majority and the minority language (see Table B.1).

Language Pattern	Successful Transfer
Min / Min	96%
Min+Maj / Min	92%
2 x Min + Maj	79%
Min / Maj	74%
Min + Maj / Maj	36%

Table B.1: Percentage of children acquiring the minority language from parents with different input patterns. Min = Minority language, Maj = Majority language, parents are separated by a forward-slash. Data from (De Houwer, 2007, p. 419)

Although De Houwer argues against a frequency based account of these findings, a rough estimate of the proportion of minority language input is revealing: By counting each language spoken by an adult as a single unit of input (so a Min+Maj/ Maj couple would have 1 Minority unit out of 3 total units), and by adding two Majority units for influence from outside the home, the proportion of Minority language input is significantly correlated with the transfer success rate ($r = 0.87$, $df = 3$, $p = 0.05$). Of course, and as De Houwer argues, the proportions of languages spoken may not be even and the amount of language input in each family may not be constant (De Houwer, 2007, p. 420).

The above study shows the importance of how language input is distributed across speakers. Typical models of language evolution (e.g. Burkett and Griffiths, 2010) model the language input as a single mass without the identity of the speaker. Even Burkett and Griffiths (2010)’s model which considers multiple teachers speaking multiple languages does not consider the distributions of languages *within* individuals. The following model looks at language evolution in the light of multiple patterns of input.

B.1.1 Stable distribution

The data from De Houwer (2007) details the probabilities of transitioning from various parental states to a state where a child is bilingual or monolingual. Unfortunately, this transition is not between equivalent states (it’s from the state of two people to the state of one), so the stable distribution over monolingualism/bilingualism given these transition probabilities is not straightforward to calculate. Furthermore, we’d have to assume that the probability of a child learning a minority language if neither parent spoke it was very near zero, leading to a stable distribution where the vast majority of the distribution was taken up by the state where both individuals speak only the majority language.

The paper does include the transition probabilities for mothers and fathers separately, however. From this we can construct the following transition matrix. We assume that all children learn the majority language. We calculate the probability of transitioning from a state where the mother only speaks the majority language to a state where the child speaks both the majority and minority languages as follows: Take the probabilities that a child will speak both languages given parental patterns where one parent only speaks the majority language (Min/Maj and Min+Maj/Maj), multiply these by the probability of this situation (provided in De Houwer, 2007) and take the sum.

Mother	Child		
	Min	Maj	Min + Maj
Min	0	5.97	94.03
Maj	0	57.23	42.77
Min + Maj	0	26.66	73.34

Table B.2: The transition probabilities (as percentages) between a mother’s language state and their child’s language state from De Houwer (2007).

This leads to a stable distribution of 38.39% speaking only the majority language and 61.60% speaking both the majority and minority languages (the proportion speaking only the minority languages is very close to zero, assuming very small probabilities for the transition to only knowing the minority language to ensure ergodicity). If we assume a chain of single individuals who adopt a hypothesis by sampling (as in Griffiths and Kalish, 2007), then this distribution also reflects

the prior bias of the learners. A more complex system (e.g. as used in Kirby et al., 2007) might suggest that the prior biases were weaker than the stable distribution. At any rate, the empirical results suggest a moderately weak bias for monolingualism.

We can attempt an analysis of the two-parent data in the following way. We'll assume there are two individuals in the population. There are 6 states: The five detailed in the tables above, plus the state of two individuals who only speak the majority language (Maj/Maj). This is the state that the system transitions to when neither child learns the minority language. The transitions probabilities to the 2 x Min + Maj state are the joint probability of two children acquiring the minority language (as given in table B.1. The transition to the Min+Maj/Maj state is the joint probability of one child acquiring the minority language and one child not acquiring the minority language. The transition to the Maj/Maj state is the joint probability of two children not acquiring the minority language. We assume that the probability of transitioning from the state where both parents only speak the majority language to that same state is equal to the transition from (2 x Min) to (2 x Min+Maj). The remainder of the transition probability is shared out equally between the other states. All other transition probabilities are set evenly so that the probabilities sum to one. This yields the transition matrix show in table B.3.

	Min/Min	Min+Maj/ Min	2 x Min+Maj	Min/Maj	Min+Maj/ Maj	Maj/Maj
Min/Min	1.00	1.00	93.93	1.00	2.99	0.09
Min + Maj / Min	2.05	2.05	87.27	2.05	6.15	0.43
2 x Min + Maj	5.50	5.50	62.69	5.50	16.49	4.33
Min/Maj	6.37	6.37	55.12	6.37	19.12	6.64
Min+Maj/Maj	7.65	7.65	12.74	7.65	22.96	41.34
Maj/Maj	0.00	5.10	0.00	0.00	0.00	100.00

Table B.3: Transition matrix estimated from De Houwer (2007)

This matrix has stable distribution shown in table B.4. In this chain, the probability that a given individual will be bilingual is 41.7%. Taking this as a rough reflection of the bias of an individual, the data suggests that individuals have a weak bias towards monolingualism.

State	Stable Distribution
2 x Min	5.10
Min + Maj / Min	5.10
2 x Min + Maj	33.75
Min/Maj	5.10
Min+Maj/Maj	10.88
Maj/Maj	40.07

Table B.4: Stable distribution (percentages) from the data from De Houwer (2007), given certain assumptions.

B.1.2 A model of language patterns

The Bayesian model described in chapter 6 was used to explore the effects of parental language patterns on the acquisition of minority languages by learners. The model was set up to contain two languages ($m = 2$). This yielded a hypothesis space of two ‘monolingual’ and one bilingual hypotheses. Several contexts were constructed to model those in De Houwer (2007), with the ‘minority’ language being defined as the language least prevalent in the input. The teachers produced data and a learner induced the most likely context. This learning in a single generation was repeated 10,000 times and the proportion of learners who ended up speaking the equivalent of the ‘minority’ language was calculated. These proportions were compared to the results from De Houwer (2007). The learning bottleneck was set to 4 utterances per teacher and the noise level was set to 0.05. The three bias parameters from the section above were used to model monolingual, uniform and bilingual prior probability distributions over the number of languages spoken by individuals.

B.1.3 Results

The tables below show the results of the modelling. Table B.5 shows the percentage of correct inductions made by learners with various biases given various input patterns. Table B.6 shows the proportions of learners who acquire the minority language. Learners with bilingual biases always acquire the minority language because they have a strong bias to select the bilingual hypothesis. Comparing this with the data from De Houwer (2007) above, only the uniform and monolingual biases fit the data. In fact, the fit is pretty close, predicting that parents who only speak the minority language transfer it with high success, whereas a learner with only one bilingual parent does much worse. However, the model over-predicts the proportion of learners who will learn the minority language in the ‘one parent, one language’ situation (min & maj). Since the ranking of the different input patterns are not correct, no further attempt to fit the model was carried out.

Context	Monolingual	Uniform	Bilingual
A & A	99.52	75.91	0.06
A & B	99.52	75.16	0.06
A & AB	3.58	64.22	2.23
AB & AB	0.17	54.4	99.22

Table B.5: Percentage of Correct Context Inductions by Learners for various input patterns.

B.2 Discussion

The results above are a crude attempt to fit the Bayesian model of bilingualism to some real data. Under some reasonable assumptions, the data fits best with a weak bias towards monolingualism. Where the model diverges from the data, this

Pattern	Real Data	Monolingual	Uniform	Bilingual
Min & Min	96	100	100	100
Min/Maj & Min	92	99.91	100	100
Min/Maj x2	79	77.0	98.28	100
Min & Maj	74	99.83	99.9	100
Min/Maj & Maj	36	51.12	88.71	100

Table B.6: Percentage of learners selecting hypotheses with the minority language, given various parental input patterns.

might suggest that additional forces are involved (that is, the Bayesian model can be seen as a ‘neutral’ model). As discussed in the main thesis, this bias might adapt to the situation the learner finds themselves in, and so assuming that all learners have the same bias might not be valid. Furthermore, this data comes from situations where there was a clear majority language. Other factors such as socio-economic status, perception of language and types of language input available could have affected these results (e.g. Scheele et al., 2010).

APPENDIX C

ITERATED REGRESSION MODEL

DYNAMICS

C.1 Measuring bilingualism

There are several ways to measure the amount of linguistic diversity in this model. The main thesis chapter uses the measure developed in the thesis of the amount of optionality that is conditioned on the social variables. This section explores some other measures.

The first measure is the number of semantic variables used in a learner's model. The second is the rank of speaker identity as a conditioning factor. This is determined by ordering the variables in a speaker's model by the magnitude with which it explains the variation in the linguistic signal (in the best model, the t-test magnitude of the probability that the coefficient of a variable is not zero). Speaker identity may not be used in the model, in which case it is assumed that it ranks lowest of all variables. Speaker identity may have different rankings for different learners in the same generation. A related measure is, in a monte-carlo simulation, the average rank of speaker identity in comparison to other variables. That is, how likely is speaker identity to become an important factor compared to other semantic variables.

There are two measures of similarity between agents in the same generation which may be thought of as addressing the similarity in internal representations (I-Language) and external expression of language (E-language, used in the main thesis chapter). The I-Language measure measures the similarity of the regression models of individuals in the model. This is calculated as the average Levenshtein-Damerau distance between the ordering of variables used in each pair of learners' models. Consider two learners, A and B. Learner A has a model that uses (ordered from most important factor to least) variables a, b, d and e . Learner B has a model that uses variables a, c, e and d . We define the distance between these two models as the edit distance: How many changes would it take to make the two models the same. To make Learner B's model the same as A's, we would have to add variable b (one step) and remove variable c (one step). We would also have to swap the orderings of d and e . Under the Damerau assumption, this swapping of adjacent symbols counts as a single step. So, A's model is 3 steps away from B's. This distance is normalised by the size of the largest model of the pair. For a population, the distance between each pair of learner's models is

calculated, and the mean is taken.

C.2 Variable selection

The stepwise regression model of cultural transmission has many parameters. While the model is flexible in some respects as a requirement of the point it is trying to make, this flexibility can make it difficult to analyse. The main method used here to explore the model dynamics is an empirical one: The model is run with many parameters and the effects are observed. A full understanding of the entire limits of the model is outside the scope of this thesis. However, some variables are not important to the main thesis, and so they will be explored briefly in order to select sensible values that can be fixed during the main exploration of the model.

This appendix proceeds by first using a monte-carlo method of exploring single generation transitions to explore possible biases in the model dynamic. Since there are few precedents for iterated stepwise regression, the exploration focuses on basic questions such as the bias for small slopes. The next section uses a simplified version of the full model (including successive generations) to confirm some of the dynamics.

Below is a table showing the parameters of the model and the selections that were made for the final model.

Model variable	Type	Choice
Semantic variables:		
Number of semantic variables (N)	Discrete, > 0	2,5
Distribution of semantic variables	normal, bimodal, uniform...	normal
Relationship between variables (y)	Continuous, 0-1	Continuous, 0-1
Transmission variables:		
Noise level	Continuous, unbounded	0.1
Sample size	Discrete, > 0	100
Stepwise regression variables:		
Allow intercept?	Yes/No	Yes, No
Stepwise regression direction	Forward/Backward/Both	Both
Information criteria (c)	continuous, unbounded	AIC,BIC
Population variables:		
Social structure integration	Set of continuous weights	Isolated, Integrated
Social structure balance	Set of continuous weights	Equal, Minority/Majority
Population size	Discrete, > 0	2,4,100
Social structure dynamics	Weight transform function	

C.2.0.1 Semantic parameter settings

The number of semantic variables will be limited to 2 and 5. Runs show that this range provides enough variation to allow interesting interactions and robust conclusions (2 variables = 3 parameters, 5 variables = 31 parameters). As described above, the distribution type of the semantic variables will be normal. The relationship between variables will be a free parameter that will be explored in the 2 variable model.

C.2.0.2 Transmission parameters

A noise level of 0.05 is enough to allow changes between generations without destroying the information. A sample size of 100 allows a stable transmission dynamics while also allowing the model to be tractable.

C.2.0.3 Regression parameters

As mentioned above, intercepts will be allowed in order to allow better fits and to reduce the bias against large slopes. The stepwise regression will be applied in both directions for a better model selection. Both AIC and BIC will be used to compare the affect of the bias against large models.

C.2.0.4 Population parameters

The population parameters are the main parameters that will be explored. They will be explained at greater length in the next section.

C.3 Monte Carlo tests

The model bias was explored by examining thousands of single-generation transitions with randomly set parameters. In all cases, there was only one teacher and one learner at each generation. The data was constructed in the following way:

1. The first semantic variable was sampled from a uniform distribution between -1 and 1.
2. The second semantic variable was created by sampling a percentage R of the first semantic variable, and a percentage $100 - R$ from a uniform distribution between -1 and 1.
3. The three model slopes (for the first semantic variable, the second semantic variable and the interaction between the two) were chosen from a uniform distribution between -1 and 1.
4. Up to two model slopes could be set to zero at random.

5. The linguistic signal was then computed from the semantic variables and the model slopes.

The learner then used this data in the same way as the full model. The change in slopes between generations was recorded.

Model variable	Value/Range
Semantic parameters:	
Number of semantic variables	2
Distribution of semantic variables	Uniform
Relationship between variables (y)	0-100%
Transmission parameters:	
Noise level	0-10
Sample size	100
Regression parameters:	
Allow intercept?	Yes,No
Stepwise regression direction	'Both'
Information criteria (c)	0, 2 (AIC), $\ln(100)$ (BIC), 10

As mentioned in the main thesis, when the slope of the teacher is small, the slope change tends to be small and random (equally distributed above and below zero). However, as the slope of the teacher increases, the slope change decreases. That is, if the teacher has a slope of 0.5, the learner change between slopes is more likely to be negative than positive (the learner has a slope smaller than the teacher). This is further illustrated by showing that, although the step sizes are normally distributed around zero, the distributions are skewed when considering generations where the teacher had a positive or negative slope separately (figure C.2). This means the model has a bias against steep slopes. This bias is mediated by the information criteria c value, R and the noise level.

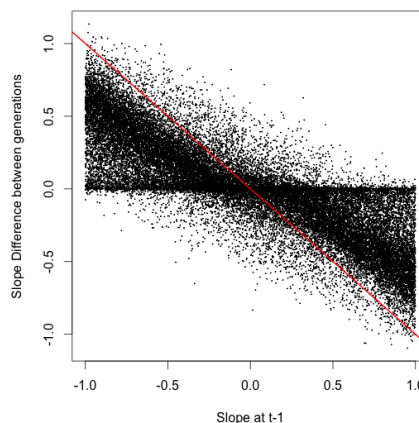


Figure C.1: The relationship between the slope of the teacher (slope at $t-1$) and the slope change between generations. The red line has an intercept of 1 and a slope of -1 on this graph, indicating a perfect (negative) 1:1 relationship between the variables.

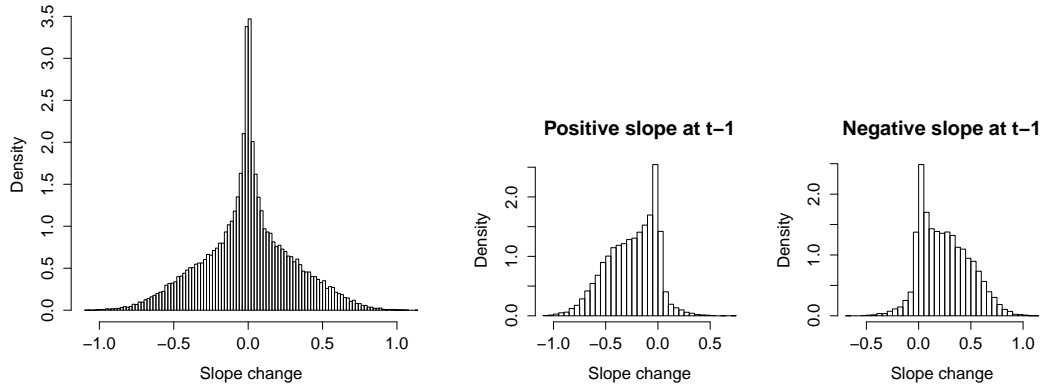


Figure C.2: Histogram of slope changes for all generations (left) and for teachers with a positive slope (middle) and a negative slope (right).

Higher values of c produce a weaker bias (figure C.3). That is, the slopes tend to move towards zero faster with lower values of c (the change between generations is likely to be greater with smaller values of c). However, if the teacher's slope is small (less than 0.2), then smaller values of c will produce a *smaller* change between generations (the regression line crosses the 1:1 line in graph C.3).

Higher values of R result in a larger bias (figure C.4). That is, when the two semantic variables are strongly linked, the slope is more likely to move towards zero.

Higher noise levels produce a stronger bias (figure C.4). That is, with a small noise level, the change in slope is not strongly related to the teacher's slope size. However, a high noise level makes it more likely that the slope will move towards zero.

Altering the stepwise regression to rule out the use of intercepts increases the strength of the bias against strong slopes. Without the intercept, the correlation between the teacher's slope and the slope change (when the teacher's slope for the semantic variable is not zero) is -0.896 (48,000 observations). With the intercept, the correlation is -0.777 (48,000 observations). For this reason, and to better fit existing linear modelling theory, the main model will include an intercept term.

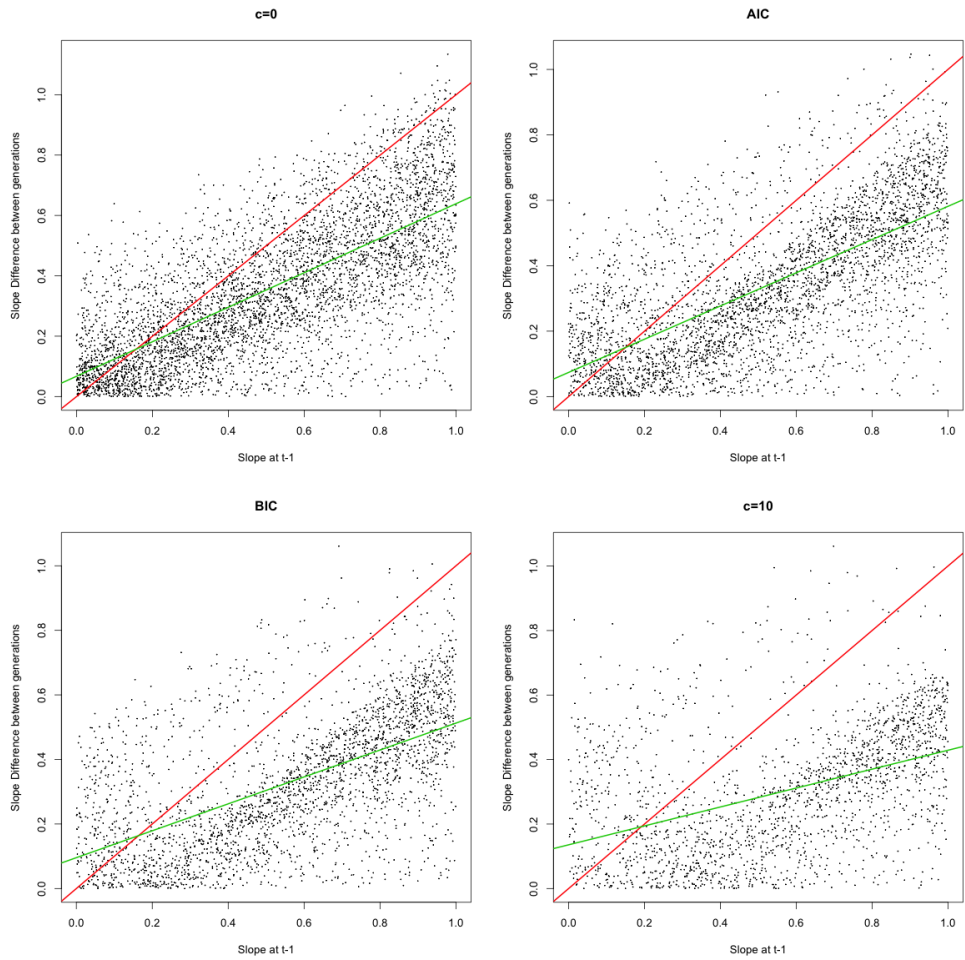


Figure C.3: The relationship between the (absolute) slope of the teacher (slope at $t-1$) and the (absolute) slope change between generations, by information criteria c value. Data shown is for semantic variables which are 100% related (although with a range of noise distorting the relationship). Red lines have an intercept of zero and a slope of 1. The green lines show the regression lines of the actual data.

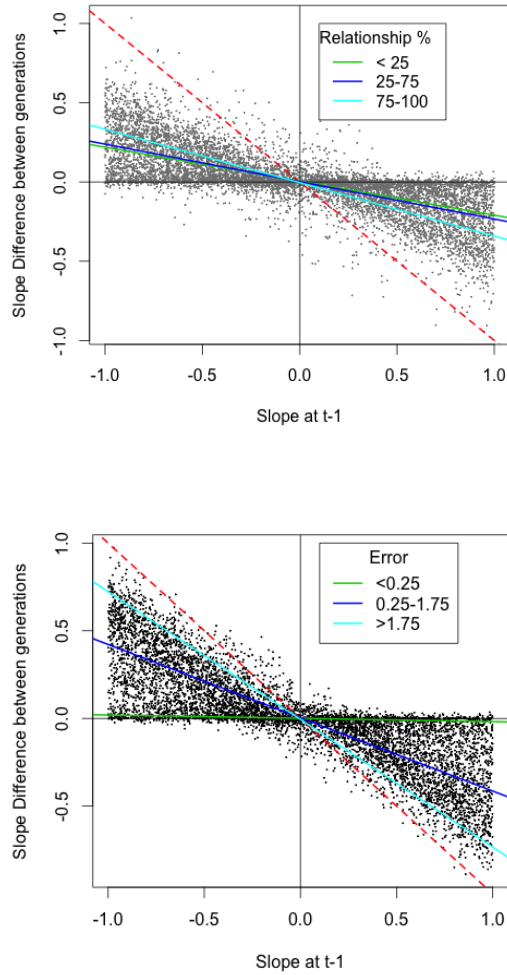


Figure C.4: The relationship between the (absolute) slope of the teacher (slope at $t-1$) and the (absolute) slope change between generations. Left: solid lines show the regression for different ranges of the amount that variables are related. Data shown is for noise levels less than one. Right: Solid lines show the regression for different ranges of noise. The dotted red line has an intercept of zero and a slope of 1.

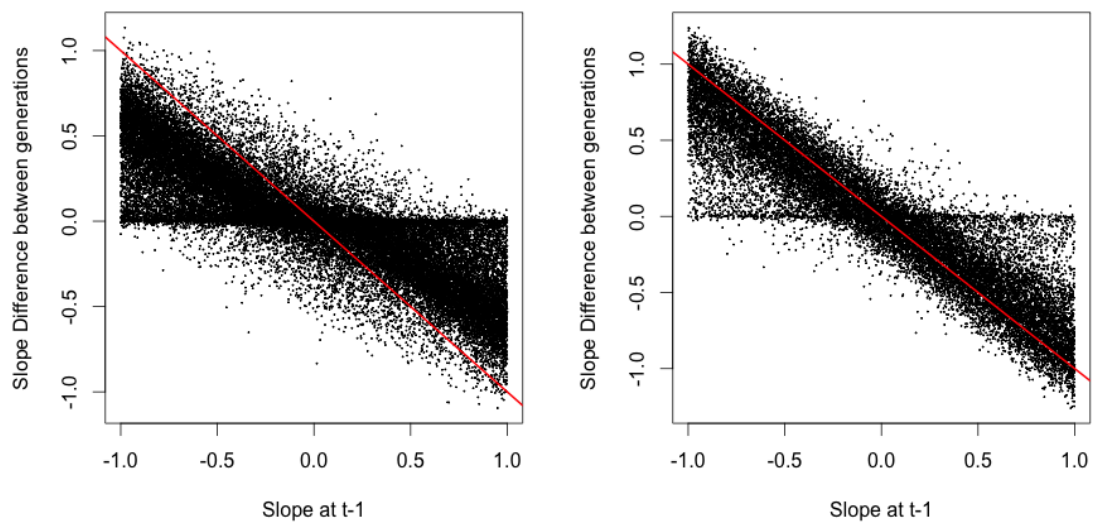


Figure C.5: The relationship between the slope of the teacher (slope at $t-1$) and the slope change between generations for models with intercepts (left) and without intercepts (right). The red line has an intercept of zero and a slope of -1 .

C.3.1 Model transitions

Figure C.6 shows the probabilities of the linguistic model transitioning from one state to another in a model with two semantic variables and the possibility of using the interaction between variables. Figure C.6 shows the stable distributions over model states given these transition probabilities. The lower the information criteria, the more variables are likely to be used in the linguistic model. Linguistic models that use the interaction term are more unstable than those that do not. If the interaction term is used, then it's likely that the other two variables will be used as well.

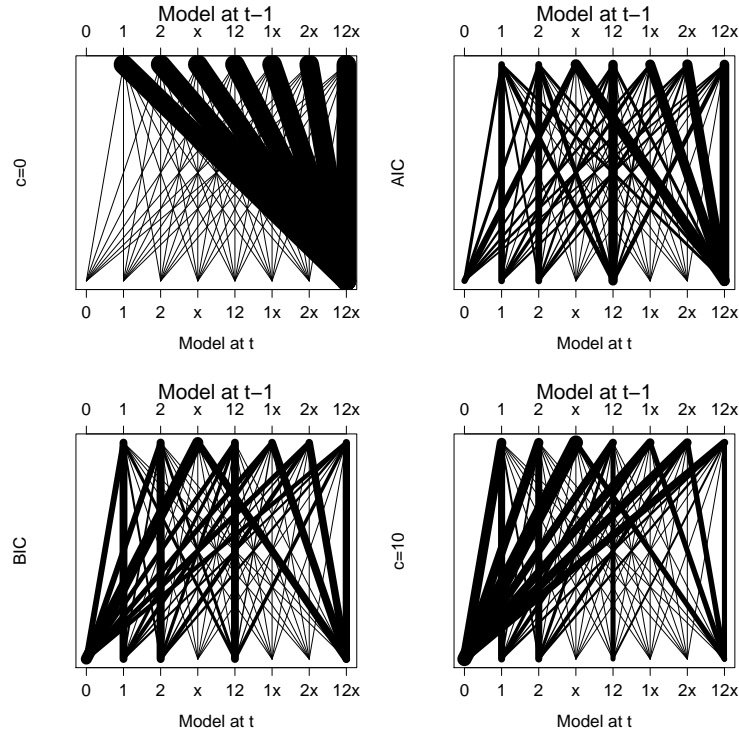


Figure C.6: Transitions between linguistic model states. The top of each graph is the state at generation $t - 1$ and the bottom is the state at generation t . Lines represent transitions, with the thickness proportional to the transition probability. 0 = no parameters, 1 = variable 1, 2 = variable 2, x = interaction between variable 1 and 2.

C.4 Model generation dynamics

This section explores the dynamics of an iterated regression model (IRM). The first section considers an IRM with a single semantic variable. This demonstrates that, if the model includes noise, the slope and intercept variables change over generations according to a random walk. The dynamics of this random walk are affected by the sample size and noise level. The distribution of the semantic variable does not affect the shape of the distribution of step sizes, but some distributions produce smaller step sizes on average.

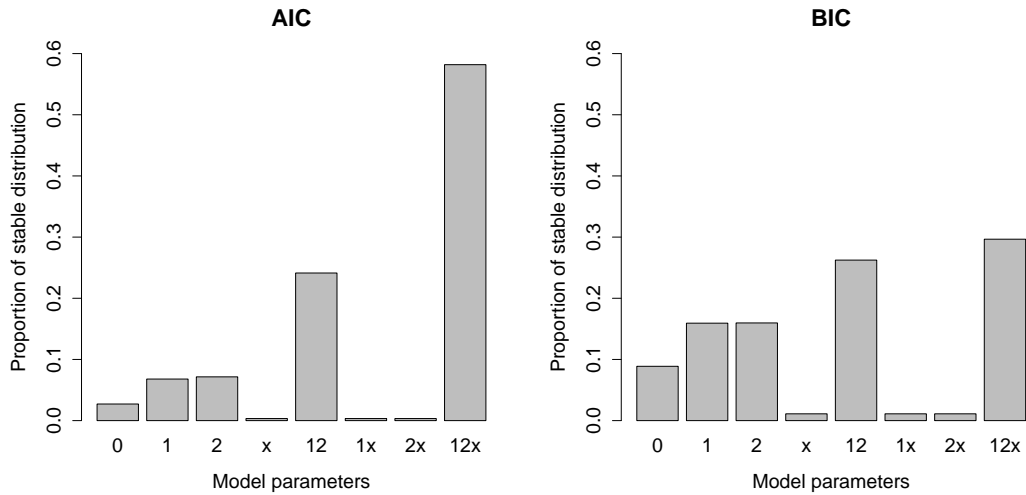


Figure C.7: Stable distributions for model types for AIC (left) and BIC (right). From first eigenvector of transition matrix for 372,600 runs. Transition probabilities were smoothed to ensure ergodicity.

The full model will include iterated multivariate stepwise regression. Before those dynamics are explored, the basic dynamics of iterated regression are examined. In this more basic model, the learner receives a single semantic variable (generated by a function) and a single linguistic variable (transmitted from previous generation). There is no exploration of models - the learner simply works out the intercept and slope that best describe a linear relationship between the two variables. The learner is then given new semantic data and produces a new linguistic signal using the linear model. This linguistic data is passed on to the next learner with a certain amount of noise. The data is assumed to come from a single speaker. The initial linguistic signal is sampled from the same function as the semantic variable.

The parameters of the model are the sample size s (how much data the learner receives and transmits), the noise level in the transmission channel e and the function that generates the semantic data.

C.5 Single semantic variable

Given a normally distributed semantic variable, the learner in the first generation will induce a model with a certain intercept and slope. With no noise, this intercept and slope will not change over time since the learner will produce a linguistic signal that perfectly corresponds to its model ¹. The only possibility of change is if the transmitted signal is generated by precisely the same semantic value, leaving no variation to analyse. In this case, however, the linear model has

¹In the actual implementation zero noise is not possible due to rounding errors.

no unique solution.

Determining the factors that affect the dynamics of the model with $e > 0$ was done empirically. 100 runs of 200 generations were carried out for noise levels 0, 0.001, 0.01 and 0.1 and sample sizes 10 to 100 in 10 sample increments.

When the noise level e is greater than zero a learner receives data whose best linear fit may deviate from the model of its teacher. The probability of change, however, is not related to the sampling size. The probability of the slope of the teacher and the slope of the learner differing is 52% for noise levels of 0.001, 0.01 and 0.1. The same probability holds for the intercept².

With transmission noise, the slope and intercept of successive generations effectively take a random walk through the space of models (see figure C.8). A random walk is a function that cumulatively adds a random amount of noise to a signal. The effect is a trajectory whose distribution above and below the mean is a bimodal distribution with size of 1 and probability of 0.5. In the case of this model, the noise being added is the difference in the models of successive generations brought about by transmission noise coupled with sampling errors.

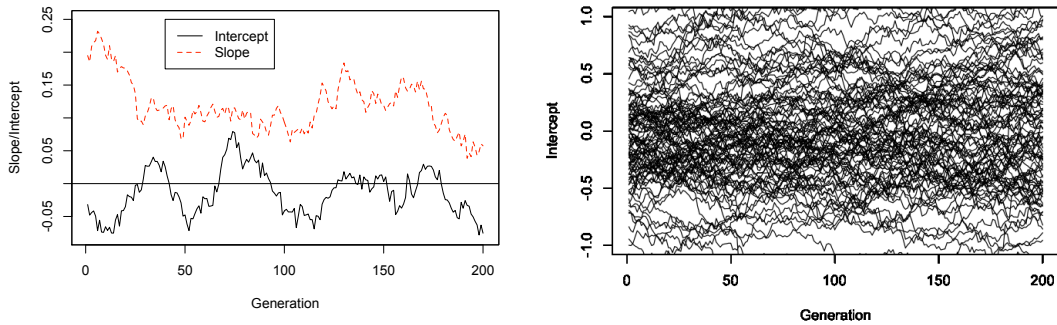


Figure C.8: The evolving slope and intercept of successive generations of learners for a single run of the model with $e = 0.1$ and $s = 50$ (left). The evolving intercept for many individual chains ($e = 0.1$, $s = 10$) (right).

The mean and standard deviation of the difference between generations is modulated by the sample size and noise level (see figures C.9-C.12). Larger sample sizes yield exponentially lower means and standard deviations. Larger noise levels yield linearly higher means and standard deviations. This is because a larger sample produces a more accurate estimate of the teacher’s model, so a decreasing sample sizes will cause exponentially larger change between generations (see also figure C.13). More noise will have the same effect, although the relationship is linear.

²This is probably affected by rounding errors. The probability of two successive models being truly the same under transmission noise is very small.

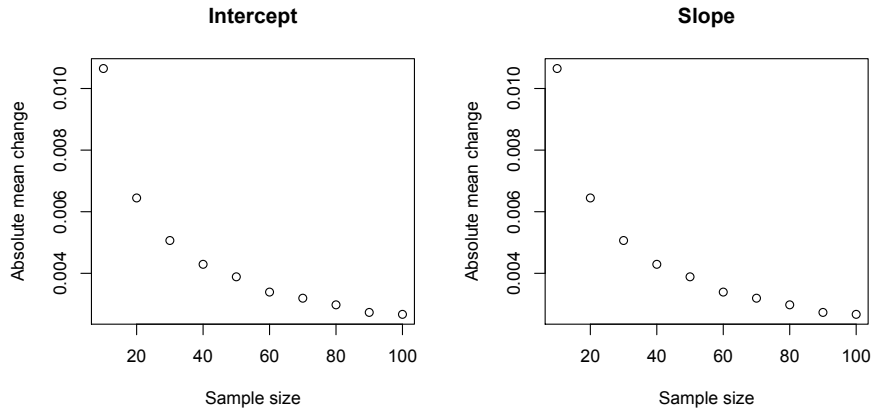


Figure C.9: The mean size of the change between generations as a function of sample size for the intercept (left) and the slope (right)

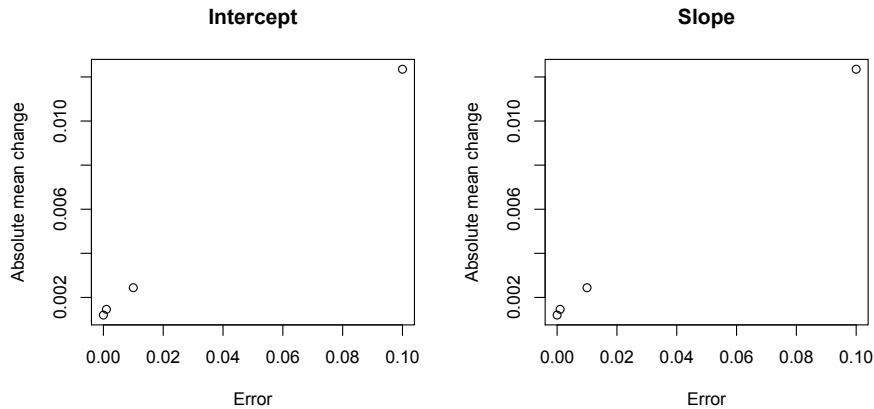


Figure C.10: The mean size of the change between generations as a function of the noise level for the intercept (left) and the slope (right)

If there was a bias, then the magnitude of the slope or intercept would correlate with the change in that variable between generations. Figure C.13 shows that this is not the case.

Despite the walk being random, the trajectory of the intercept or slope can be highly positive or negative. This can be measured by the correlation between the intercept or slope value and the generation number. A large correlation coefficient suggests a steadily increasing or decreasing variable. Figures C.14 and C.14 show the distribution of correlation coefficients for intercept and generation. The results of the model do not seem to differ from a true random walk process with the same parameters.

A random walk is a stochastic process which can be thought of as a walker taking a number of independent identically distributed steps (see textbooks on

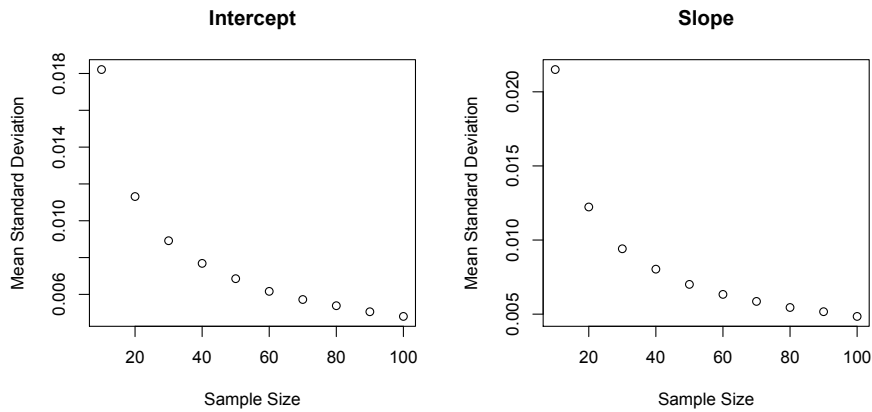


Figure C.11: The mean standard deviation of the change between generations as a function of sample size for the intercept (left) and the slope (right)

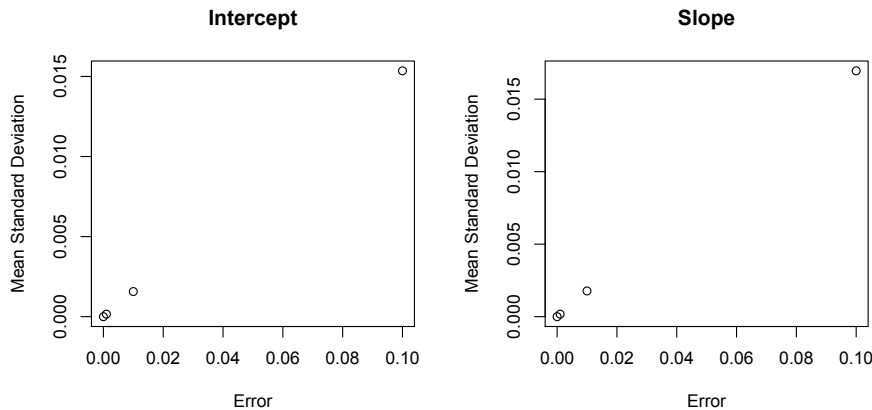


Figure C.12: The mean standard deviation of the change between generations as a function of the noise level for the intercept (left) and the slope (right)

complexity, e.g. Mainzer, 2004). The walker takes n steps of size x_i over time t . The position of the walker is the sum of the steps $S_n = x_1 + x_2 + \dots$. The variance of the stochastic process $x(t)$ grows linearly with the number of steps. The shape of the probability density function of the steps $P(x_i)$ affects the shape of the probability density function of positions $P(S_n)$.

However, the probability density function of the semantic variables does not affect the shape of the probability density function of the slope and intercept variables. Compare the graph showing the probability density function for translation for a true random walk and for the IRM model (figure C.16). The density for the IRM is normal with a high kurtosis with the semantic variable being drawn from a normal, uniform or bimodal distribution. True random walks look slightly different, with the bimodal distribution looking more triangular. The distribution of step sizes is also similar for models with different semantic variable distributions (figure C.17). However, the kurtosis is different: A normally distributed semantic

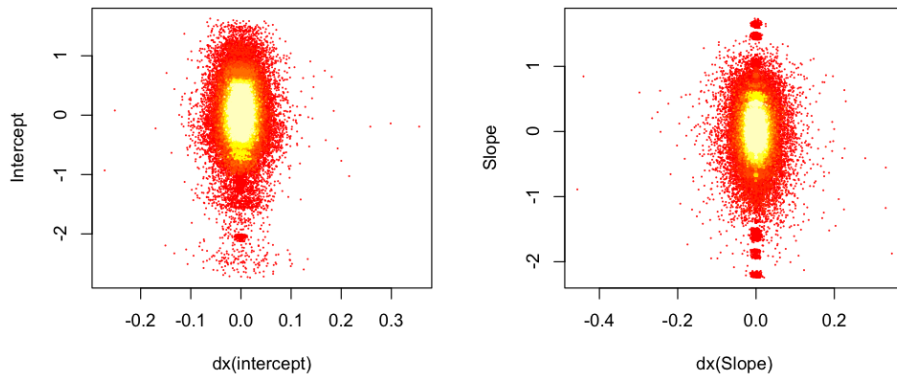


Figure C.13: The intercept (left) and slope (right) as a function of the change in intercept and slope between generations for many runs and parameterisation of the model. The sample size is indicated by colour with lower sample sizes being coloured red and higher sample sizes being coloured yellow and white.

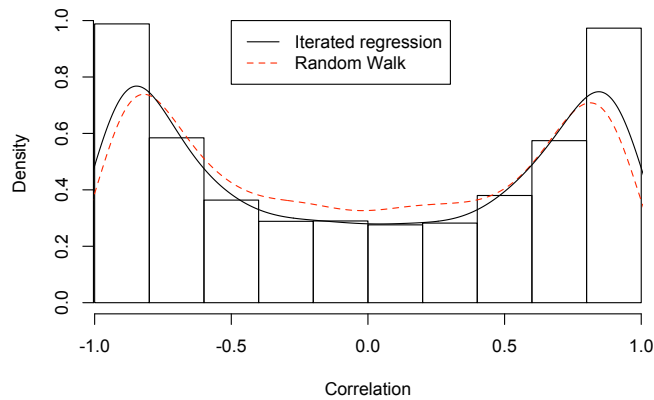


Figure C.14: The distribution of correlation coefficients for intercept and generation. A high positive correlation indicates a steadily increasing intercept value. The red dotted line shows the distribution for a simulated random walk with the same parameters and number of runs.

variable produces a step-size distribution with the lowest kurtosis (34 in the sample of non-zero noise level model runs), followed by the uniform distribution (54) and the bimodal distribution (531). That is, a bimodal semantic distribution tends to produce lower step-sizes than a normal distribution on average. The process of linear regression neutralises the differences in semantic variable distribution to produce step sizes that are normally distributed (although perhaps the distributions are closer to a Cauchy distribution).

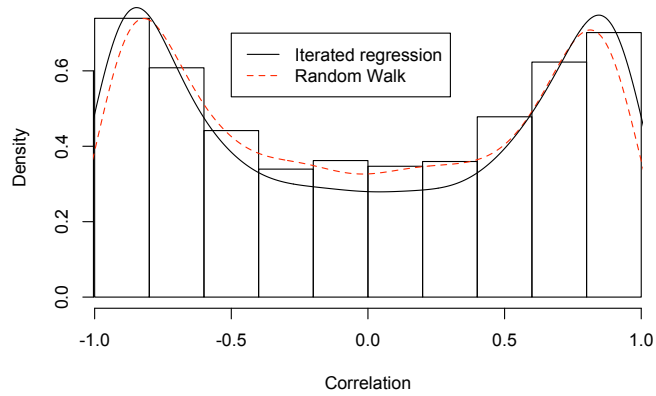


Figure C.15: The distribution of correlation coefficients for slope and generation. A high positive correlation indicates a steadily increasing slope value. The red dotted line shows the distribution for a simulated random walk with the same parameters and number of runs.

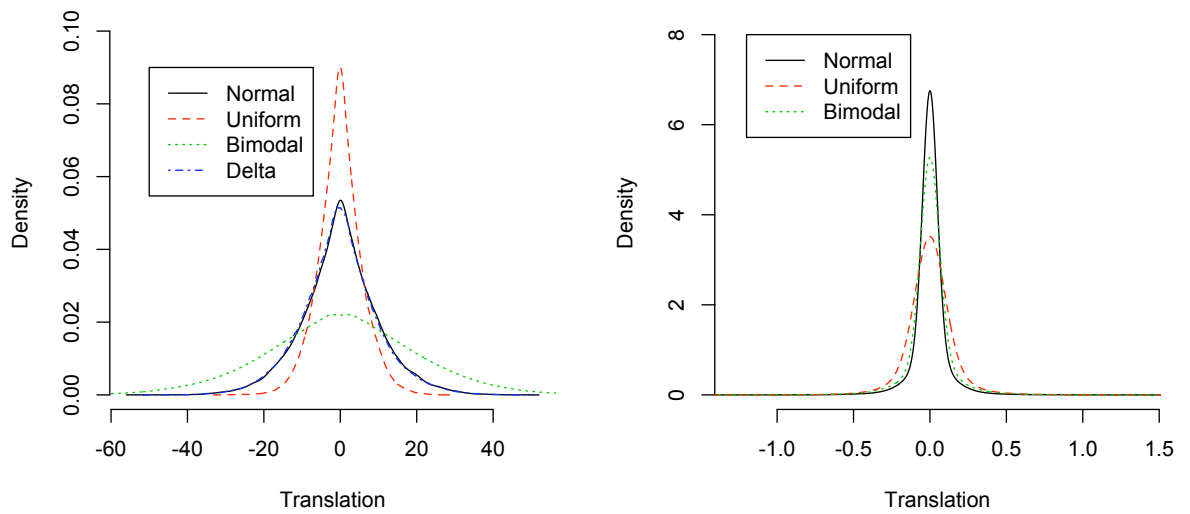


Figure C.16: Left: Probability density function of the translation from the origin for a simulated random walk with step sizes sampled from a normal distribution, uniform distribution, bimodal distribution and delta function (the normal and delta lines lie on top of each other). Right: The probability density function of translation from the origin for the intercept variable for the simple IRM model with a semantic variable distributed normally, uniformly and bimodally.

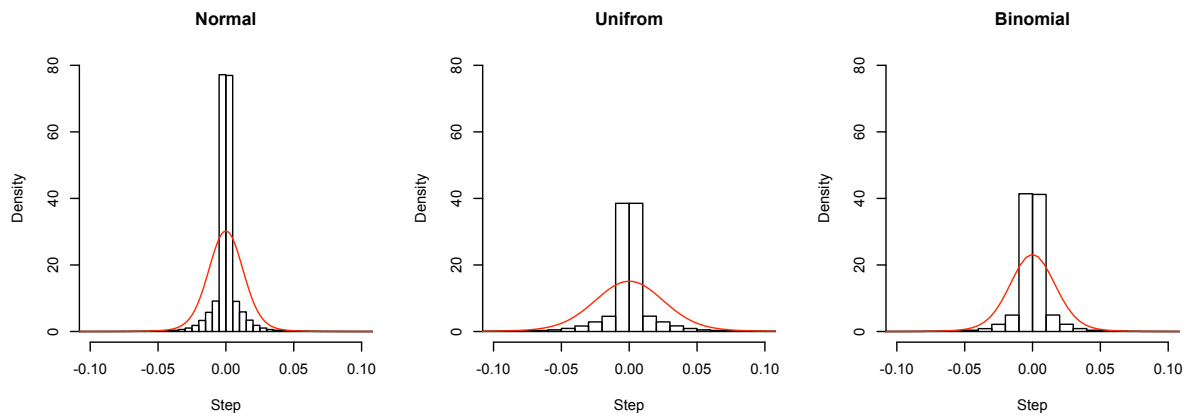


Figure C.17: The probability density function of step sizes in the intercept variable for a semantic variable with a normal distribution (left), uniform distribution (middle) and bimodal distribution (right). The lines show kernel density estimates (gaussian kernel estimator with a bandwidth of 0.02).

C.6 Multiple semantic variables: Identical distributions

Including more than one semantic variable changes the model qualitatively. First, there is the stepwise regression element, whereby variables can be excluded from the model. There is also the possibility of including interactions between variables.

The model was run with two semantic variables, each being drawn from a normal distribution with a mean of zero and a standard deviation of 1.

C.6.1 Generation

C.6.1.1 Number of variables in the model

The average number of variables in the model increases over generations (figure C.18).

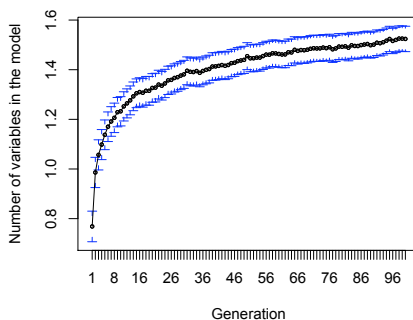


Figure C.18: The mean number of variables in the model for runs with a noise level greater than zero (with 99.99% confidence intervals).

C.6.1.2 Intercept

The intercept rapidly converges for small noise levels, but increases slowly for larger noise levels (figure C.19). The significance of the intercept is not affected by the information criteria (figure C.20).

C.6.1.3 Slope

The slope magnitude starts low, increases to a high level in the first generation and then decreases over generations to a more stable mean which is dependent on

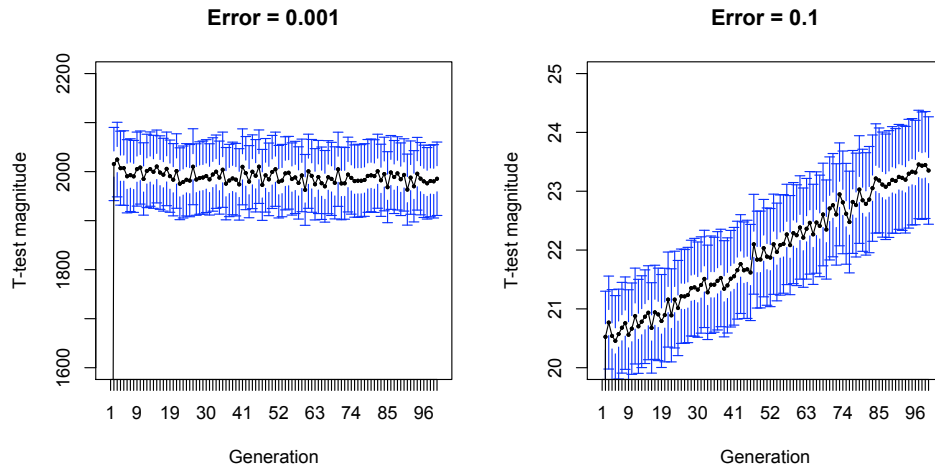


Figure C.19: The mean significance of the intercept in models over generations for different error values.

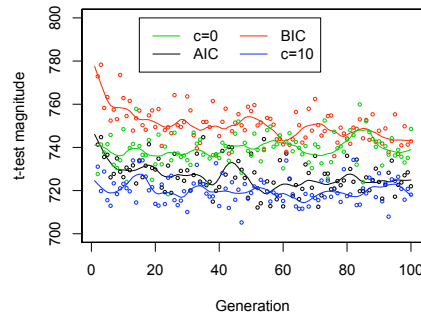


Figure C.20: The significance of the intercept over generations for different information criteria.

the noise level (figure C.21) and the information criterion (figure C.22). The slope significance (as measured by the t-test magnitude) follows a similar pattern (see same figures). As the mean number of variables increases over generations, there is more competition between variables on average which means relatively lower slopes and slope significance in the model. This is demonstrated by observing models where the information criterion variable c is zero (i.e. there is no punishment for including a variable in the model). The slope and slope significance in these models stays relatively stable over generations (unlike the for other values of c) because both variables are being included in learner's models from the start (see figure C.34 in section C.6.4).

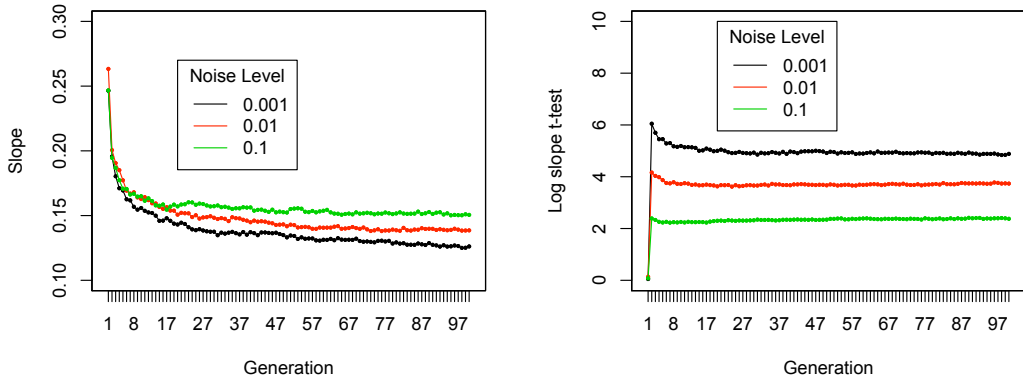


Figure C.21: The mean slope magnitude (left) and slope significance (right) over generations by noise level.

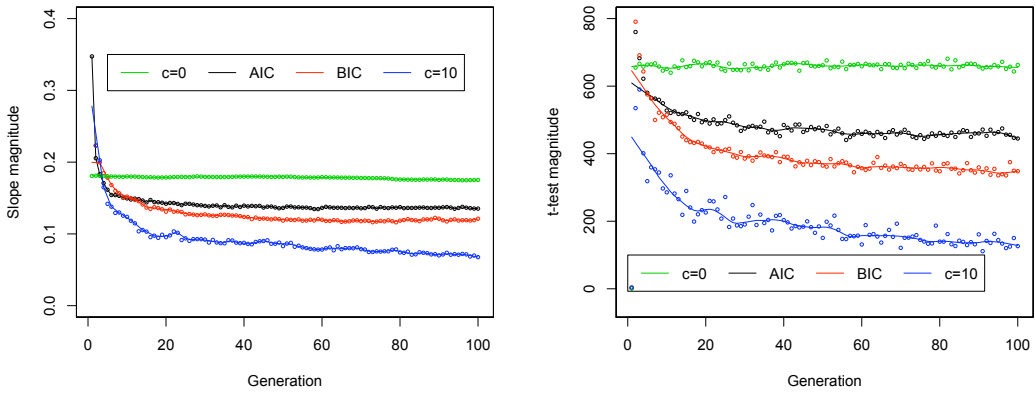


Figure C.22: The mean slope magnitude (left) and slope significance (right) over generations with different information criteria.

C.6.1.4 Slopes ratio

Slopes in a model with two semantic variables can be compared using a ratio. This is calculated as the smallest slope magnitude divided by the largest slope magnitude. For instance, a model with one slope with a value of 0.2 and another of 0.4 would have a ratio of 0.5. The smaller the ratio, the more similar the two slopes are.

Although the mean slope ratio has a wide range over generations, the variance in slope ratios over many parameterisation of the model does decrease marginally in the first 50 generations (figure C.23). However, the trajectory of the slopes ratio for individual runs can be quite chaotic or hardly moving (figure C.24). That is, the slopes ratios for some runs with the same parameterisation have a large standard deviation, while others have a very small standard deviation. This sug-

gests that there are two types of model dynamic (a bifurcation) divided by initial conditions. Note that while many model runs with a low slope ratio standard deviation have a low slope ratio, this isn't always the case.

The most crucial factor that decides the model dynamic is the number of variables in the first generation's model (figure C.25). That is, the model dynamics are sensitive to initial conditions (the number of variables in the model in the first generation is largely an accident of how the first linguistic variable is distributed). If the first model has no semantic variables, then the standard deviation of the slope ratio (how much the balance between the two semantic variables changes) will be considerable. If there are one or two semantic variables in the first generation's model, however, the standard deviation of slopes ratios will be small (i.e. there is no drastic change in how the semantic variables are used by learners over generations). Note that the relationship has weakened after 50 generations (figure C.25, right). This suggests that a change in the distribution of semantic variables can make a qualitative difference to the dynamic of the model.

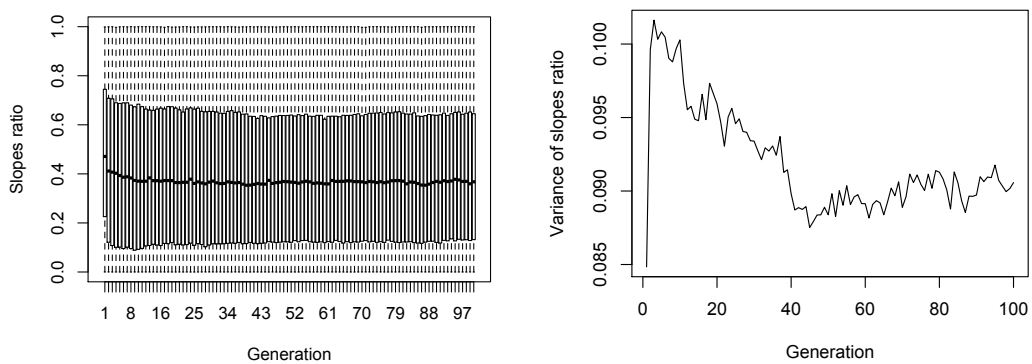


Figure C.23: The mean slope ratio (box-and-whisker plot, top) and variance of slope ratios over generations (bottom).

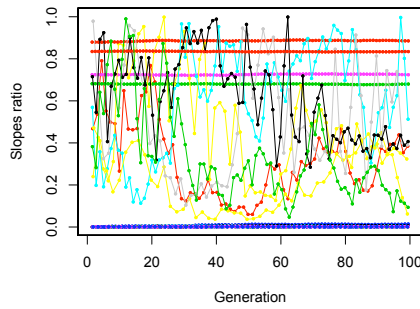


Figure C.24: The slopes ratio over generations for 20 runs of the model with a noise level of 0.001, a sample size of 80 and AIC model selection. Some trajectories are chaotic while others change only very little.

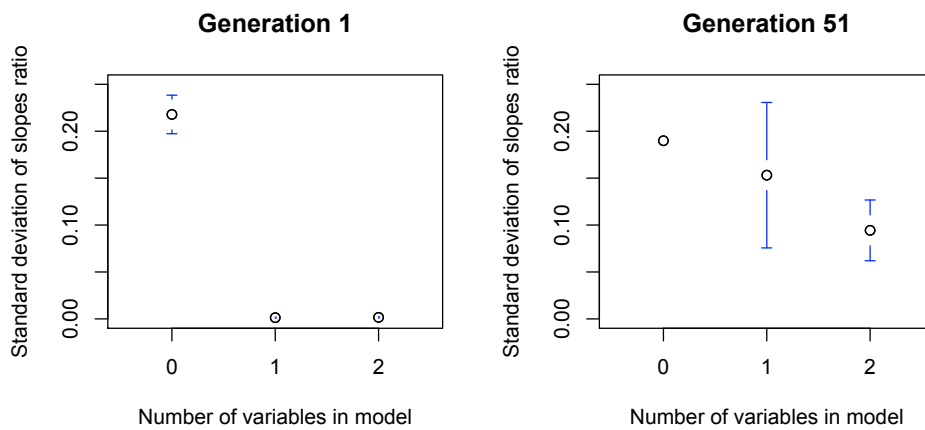


Figure C.25: The standard deviation of slopes ratio (how the balance shifts between semantic variables over generations) as a function of the number of semantic variables in the model at the first generation (left) and after 50 generations (right).

C.6.2 Noise level

C.6.2.1 Number of variables in the model

Overall, most models include 2 variables. There is a small affect of noise on the number of semantic variables in the model: With zero error, the proportion of generations with to semantic variables is greater.

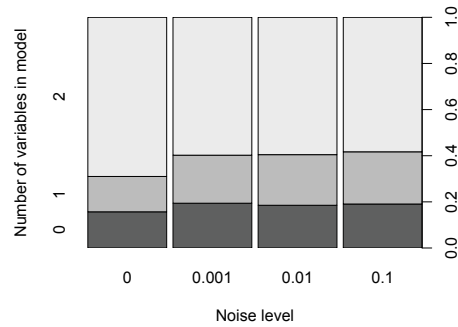


Figure C.26: The number of semantic variables used in the model as a function of the noise level.

C.6.2.2 Intercept

The noise level is not related to the mean intercept (see figure C.27). However, the noise level does affect the significance of the intercept in the model, with greater noise levels leading to a lower intercept significance (figure C.27).

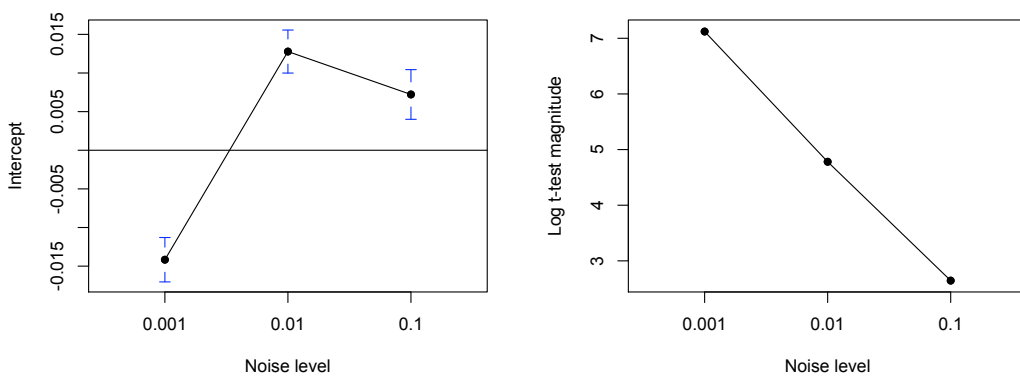


Figure C.27: The mean intercept (left) and intercept significance (right) as a function of noise level.

C.6.2.3 Slope

A higher noise level produces larger slopes on average (figure C.28). This is because a higher noise level increases the average step size in the random walk. An increase in noise level from a small amount of noise has a bigger effect on the slope than the same increase from a large amount of noise.

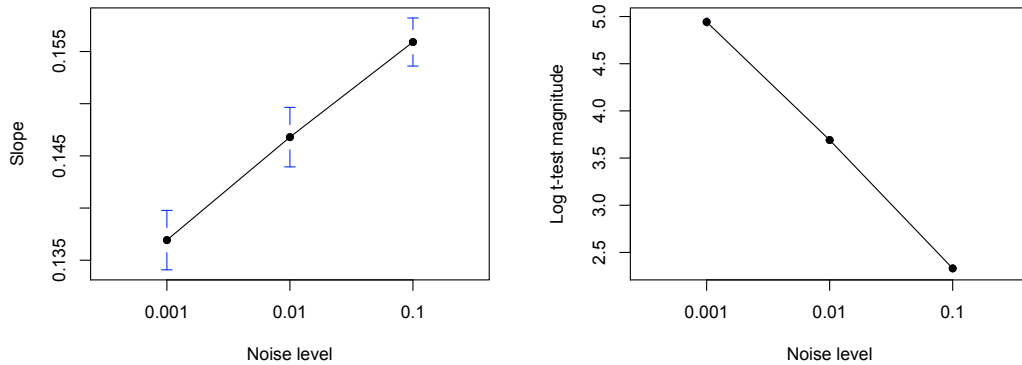


Figure C.28: The slope magnitude (left) and slope significance (right) as a function of noise level.

C.6.2.4 Slopes ratio

The ratio between the slopes of the two semantic variables (when there are two variables in the model) decreases linearly as the noise level increases (figure C.29). That is, with less noise, the slope for one variable will be many times greater than another. Under uncertainty (more noise), the effect of a single variable decreases and the slopes become more similar. The relationship between the slope significance ratio and noise level is very similar because of the strong relationship between the ratios of slopes and slope significance (see section C.6.5).

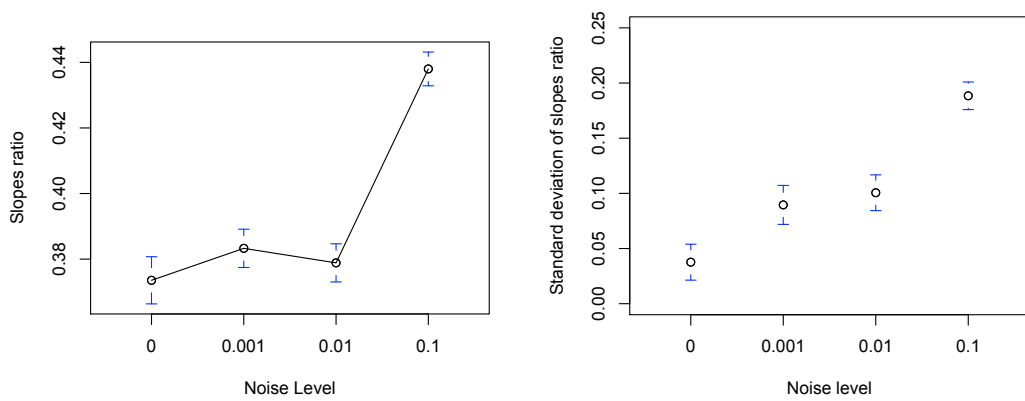


Figure C.29: The ratio between the magnitude of the slopes of semantic variables (left) and the standard deviation of the slopes ratio (right) as a function of noise level.

C.6.3 Sample Size

C.6.3.1 Number of variables in the model

The sample size has a small effect on the number of variables used in the model, with smaller sample sizes resulting in a greater proportion of two-variable models (figure C.30).

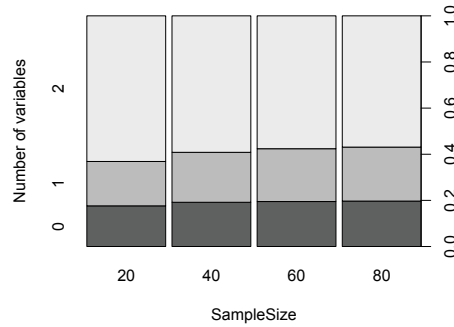


Figure C.30: The slope magnitude as a function of noise level.

C.6.3.2 Intercept

Larger sample sizes produce smaller intercepts on average (figure C.31). This is because a larger sample size produces a more accurate estimate of the semantic variables, so the model can rely less on the intercept to explain the variance.

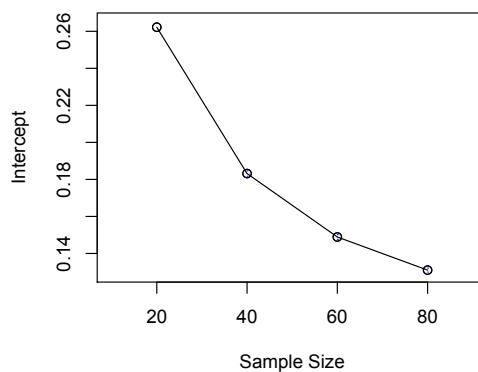


Figure C.31: The intercept magnitude as a function of the sample size for runs with a noise level greater than zero.

C.6.3.3 Slope

Larger sample sizes produce smaller slopes on average (figure C.32, left). There is no strong relationship between slope significance and sample size (figure C.32, right). The pattern in figure C.32 is different for sub-samples with different noise levels. That is, sample size has a much weaker effect on the slope significance than noise level (compare with figure C.28).

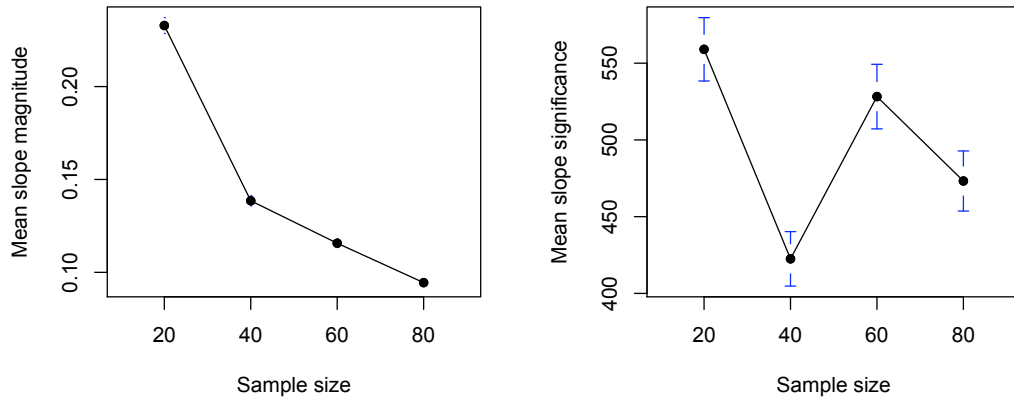


Figure C.32: The slope magnitude (left) and slope significance (right) as a function of the sample size for runs with a noise level greater than zero.

C.6.3.4 Slope ratio

There is no relationship between the sample size and the ratio between slopes in a two-variable model (although the graph suggests a u-shaped relationship, the pattern is different for sub-samples with different noise levels), nor does sample size affect the standard deviation of the ratio between slopes over generations of the same run (figure C.33).

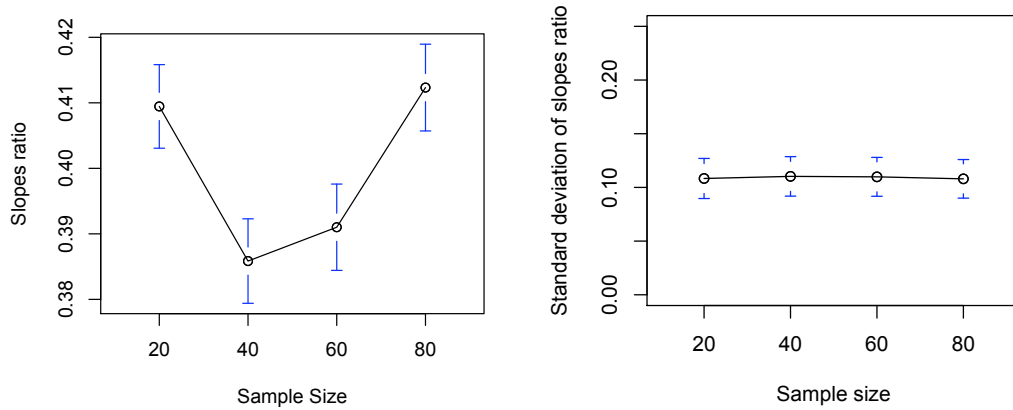


Figure C.33: The ratio between slopes in a two-variable model (left) and the standard deviation of the slopes ratio within runs (right) as a function of the sample size for runs with a noise level greater than zero.

C.6.4 Information criteria

C.6.4.1 Number of variables in the model

The number of variables in the model is affected by the parameters of the information criteria by which the stepwise regression is evaluated. Low values of c (how much the stepwise regression punishes models with greater number of variables, see section A.1) result in a greater proportion of the models that use both semantic variables (figure C.34). Higher values of c result in a greater proportion of models with no semantic variables.

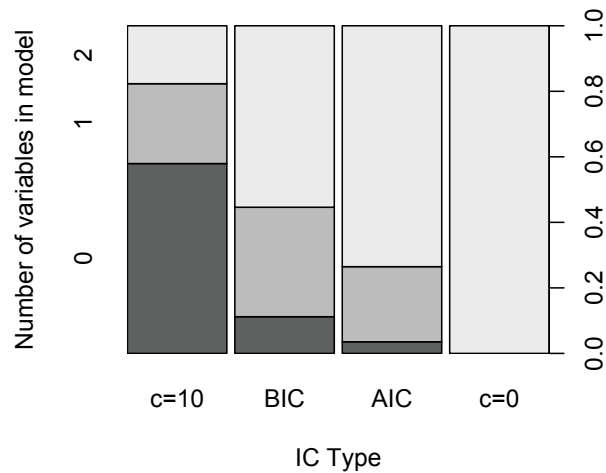


Figure C.34: The proportion of models using 0, 1 and 2 variables for stepwise regressions with different information criteria.

C.6.4.2 Intercept

The intercept is lower on average for higher values of c (figure C.35).

C.6.4.3 Slope

Higher values of c produce lower slopes on average (figure C.35).

C.6.4.4 Slopes ratio

Extreme values of c (very low or very high) result in higher ratios between slopes in models with two semantic variables (figure C.36). AIC and BIC result in

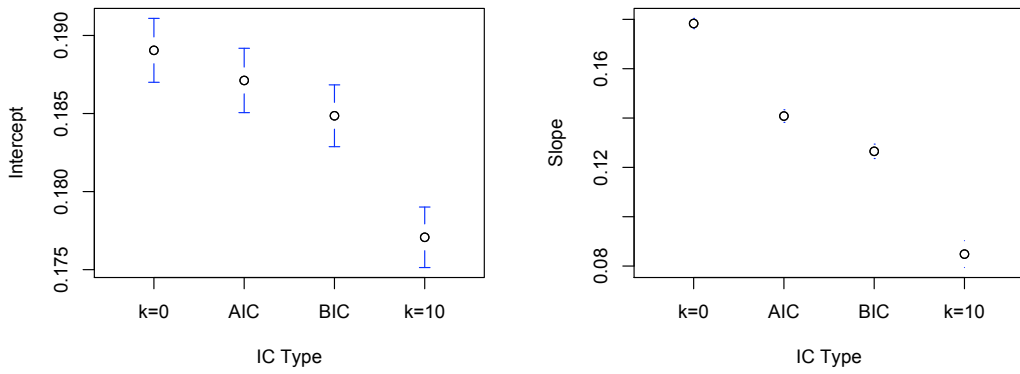


Figure C.35: The mean intercept (left) and slope (right) magnitude for stepwise regressions with different information criteria.

models where one variable has a much higher slope than the other.

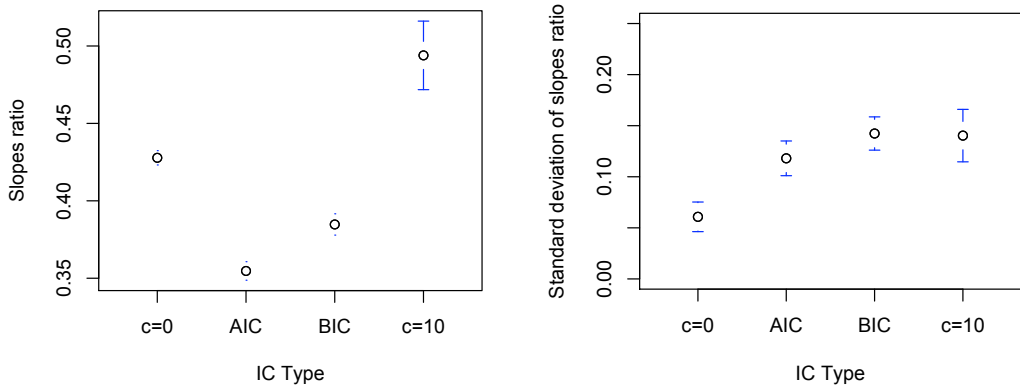


Figure C.36: The mean ratio between slope magnitudes (left) and standard deviation for the ratio between slopes within runs (right) for models with two semantic variables for stepwise regressions with different information criteria.

C.6.5 Relationship between slopes ratio and slope significance

Although there is only a weak relationship between the magnitude of a semantic variable's slope and the significance of the variable in the model (as measured by its t-test value, see figure C.37), in models with two semantic variables there is a positive linear, heteroscedastic relationship between the ratio of slope magnitudes and ratio of slope significance (figure C.38).

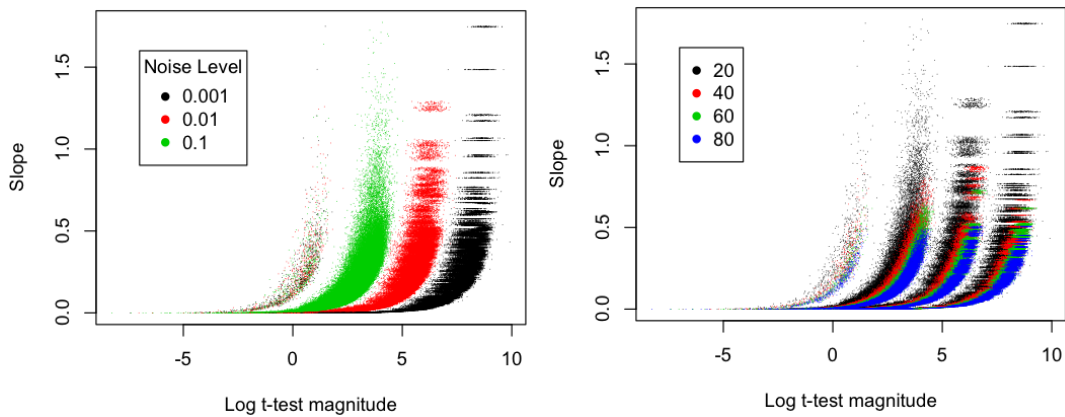


Figure C.37: Relationship between slope magnitude and slope significance for several parameterisations of the model (noise level is greater than zero) coloured by noise level (left) and sample size (right). Each point represents an individual's linguistic model in a particular generation.

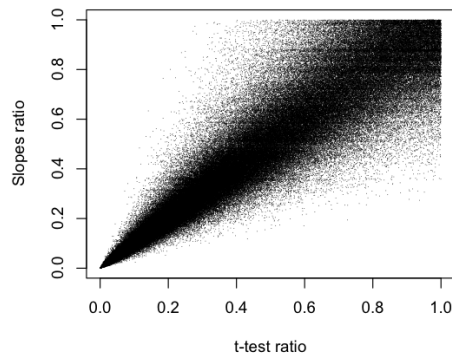


Figure C.38: Relationship in models with two semantic variables between the ratio of slope magnitudes and ratio of slope significance for several parameterisations of the model (noise level is greater than zero). Each point represents an individual model in a particular generation.

C.7 Multiple semantic variables: Different distributions

The model was run with two semantic variables sampled from different normal distributions. Variable 1 had a mean of -1 and a standard deviation of 0.5, while variable 2 had a mean of 1 and a standard deviation of 1.5. The linguistic variable was initialised with a mean of 0 and a standard deviation of 1.

The model prefers variables with lower variance, assigning them higher slopes which account for more of the variance (figure C.39).

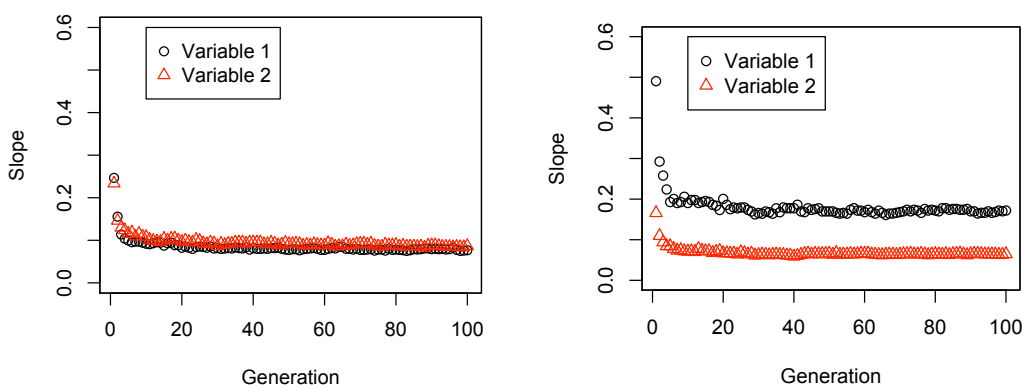


Figure C.39: Mean slope magnitudes for variable 1 and variable 2 by generation for 100 runs of models using AIC with a noise level of 0.001 and a sample size of 80. Left: Semantic variables with the same distribution (mean = 0, sd = 1). Right: Semantic variables with different distributions (Variable 1: mean = -1, sd = 0.5; Variable 2: mean = 1, sd = 1.5).

C.7.1 Noise level

C.7.1.1 Slopes ratio

The pattern for slopes ratio and information criteria type is broadly the same: Extreme values of c produce higher slope ratios on average (figure C.40). However, with differently distributed variables, the $c = 0$ (no punishment of models with many variables) setting now produces higher slopes ratios than the $c = 10$ setting (harsh punishment of models with many variables).

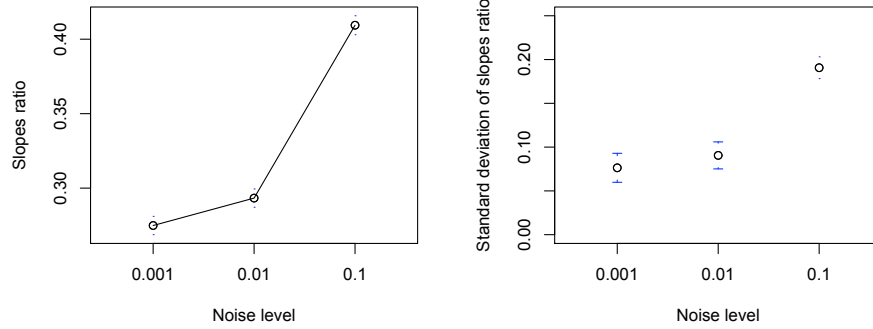


Figure C.40: The ratio between slopes (left) and standard deviation for the ratio between slopes within runs (right) for semantic variables in two-variable models as a function of the noise level.

C.7.2 Information criteria

C.7.2.1 Number of variables in the model

With high levels of c , differently distributed variables produce more single-variable models with differently distributed variables than identically distributed variables (figure C.41).

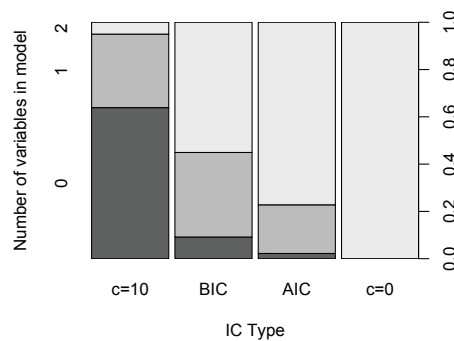


Figure C.41: The proportion of models using 0, 1 and 2 semantic variables by information criteria.

C.7.2.2 Intercept

The significance of the intercept term tends to start high and decrease to a plateau over generations. The eventual stable significance of the intercept interacts with the information criteria (figure C.42). While there is no change for low values of c (or a rapid convergence to the plateau), higher values decline over about 50 generations. The intercept in models using BIC begin by explaining more of the variance on average than models using AIC, but by about 50 generations this pattern has reversed. After the plateau, more extreme values of c result in

higher significance for the intercept on average. This contrasts with the results for a model with identically distributed semantic variables where the intercept significance is less affected by the information criterion parameter (see figure C.20).

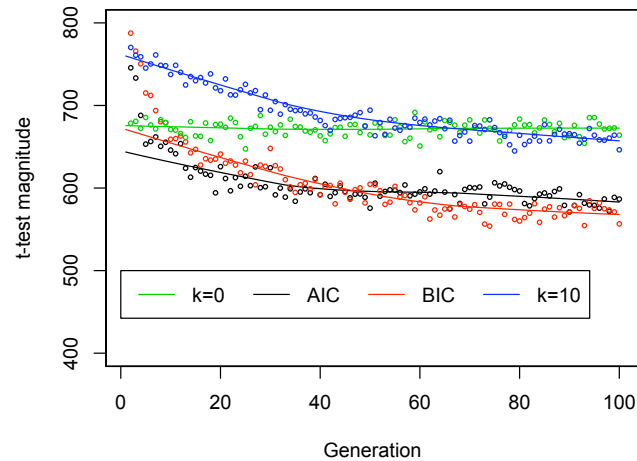


Figure C.42: The significance of the intercept over generations by information criteria for a model with differently distributed semantic variables.

C.7.2.3 Slope

Despite the difference above, there is no difference between the identically distributed and differently distributed semantic variable models in how the information criterion affects the slopes.

C.7.2.4 Slopes ratio

AIC tends to lead to higher slopes ratios and higher standard deviations of slopes ratios than BIC (see figure C.43).

C.8 Allowing interactions

The probability of an interaction between variables increases over generations. However, this just reflects the increasing average number of variables in the model (see section C.6.1). For models with two variables, the probability of an interaction between them does not increase significantly.

The presence of an interaction is sensitive to initial conditions. The following statistics are for 4800 runs with various parameter settings and a noise level

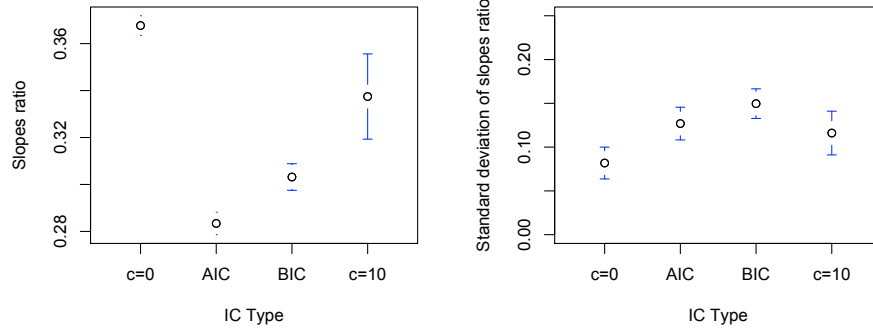


Figure C.43: The ratio between slope magnitudes (left) and the standard deviation for the ration between slope magnitudes within runs as a function of information criteria.

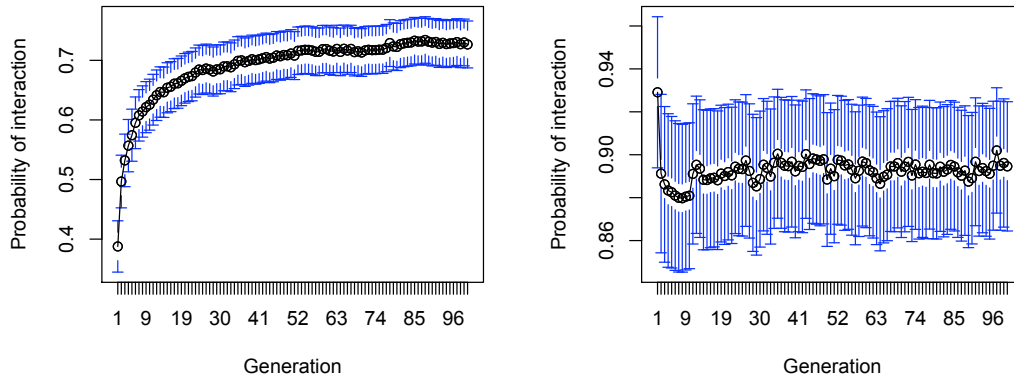


Figure C.44: The proportion of generations with models with an interaction between the two semantic variables over generations for all models (left) and models with two semantic variables (right).

above zero. 36% of runs which had an interaction in the first generation maintained the interaction for all subsequent generations. 1% of runs which started with no interaction maintained no interaction for all subsequent generations. 99% of generations following a first generation with an interaction had an interaction, while only 50% of generations following a first generation with an interaction had an interaction.

For models with two variables, the noise level has a small effect on the probability of an interaction (more noise decreases the probability of an interaction), while the sample size only has a marginal effect (a larger sample size decreases the probability of an interaction, figure C.45). The probability of an interaction in models with two variables decreases linearly with the information criteria c value (figure C.46). This just reflects the extra punishment for including an interaction

term.

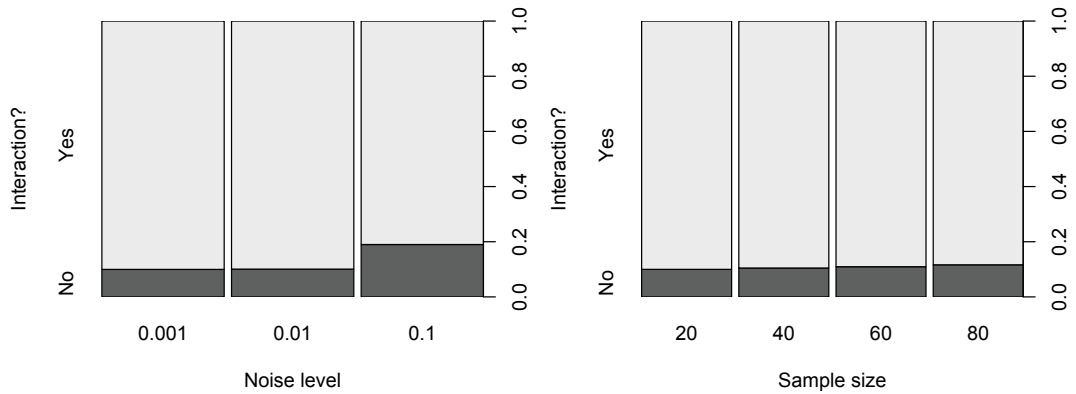


Figure C.45: The proportion of generations with models with an interaction between the two semantic variables as a function of noise level (left) and sample size (right).

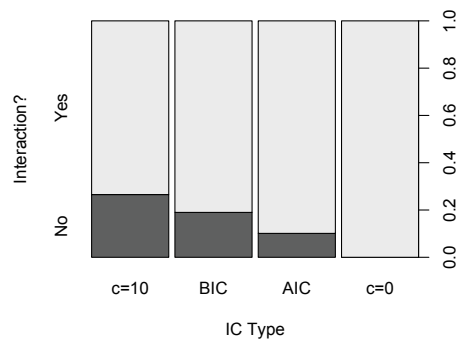


Figure C.46: The proportion of models with two semantic variables with an interaction as a function of information criteria type.

C.9 Full model examples

The model was run with the dynamic community structure described in chapter 7. Two communities of two individuals each went through stages of being isolated and integrated. The factors that affected the model dynamics were investigated by running many simulations and using a linear regression to determine important factors. The following measures were used:

- Speaker ID rank mean: The mean speaker ID rank for all linguistic models for all individuals.

- Speaker ID rank sd: The standard deviation for speaker ID rank for all linguistic models for all individuals.
- Speaker ID rank ratio: The ratio between communities of their mean rank of speaker ID in their linguistic models. A ratio of 1 indicates that communities have a similar speaker ID rank. Deviations from 1 indicate different ranks. The comparison is always community 1:community 2, so a value above 1 means that speaker ID has a higher mean rank in community 1 (speaker ID is less important). Every ratio in this set works in this way.
- Linguistic output mean ratio: The ratio between communities of their linguistic signal mean values. That is, for each community, calculate the mean of all linguistic signals produced, then compare the means for both communities.
- Linguistic output sd ratio: The ratio between communities of their linguistic signal standard deviations.
- Linguistic output mean diff: The absolute difference between each community's mean linguistic signal.
- Model fit ratio: How well the linguistic models in a community fit the linguistic input they receive. Calculate how well each speaker's model fits the data they observe (using the sum of squares of the residuals), take the average within each community and compare these values. One might predict that a community with a better fitting model would change less during contact.
- Model similarity ratio: A measure of the similarity between linguistic models in a community. Within each community, compare how well each individual's model captures the data that another observed, then compare these values between communities. This is calculated by comparing the linguistic output of each individual given the same semantic data, then taking the mean of the residuals. One might predict that the superstrate community would be the one with the greatest similarity between individuals.
- Model size ratio: The ratio between the mean number of parameters in the linguistic models of each community. One might predict that the community with the largest model would change less.
- Interactions ratio: The ratio between the mean number of interaction parameters in the linguistic models of each community.
- Model size mean: The mean number of parameters in all linguistic models.
- Transition generation: The number of generations after the community structure transition has started.
- Generation: The number of generations since the simulation started.

For looking at positive bilingualism situations where there is an imbalance in the extent to which each community adapts to the other, a measure of the extent to which a community’s linguistic signal changed over generations was used. The comprehensive intelligibility score, as described in section 7.3.2 of chapter 7 was calculated between individuals within the same community but belonging to different generations. The mean for each community was taken. The community with the highest value can be identified as the ‘superstrate’ community, since its linguistic signal has changed the least.

C.9.1 Explaining bilingualism scores

Table C.1 shows the results of a linear regression fit to the data from 100 simulations with the dynamic social structure described in chapter 7. The model explained 16.9% of the variation in the bilingualism scores (adjusted R^2 , $F(30,7368)=51.15$, $p < 0.000001$). See section 7.7.1 on page 146 for a description of these results.

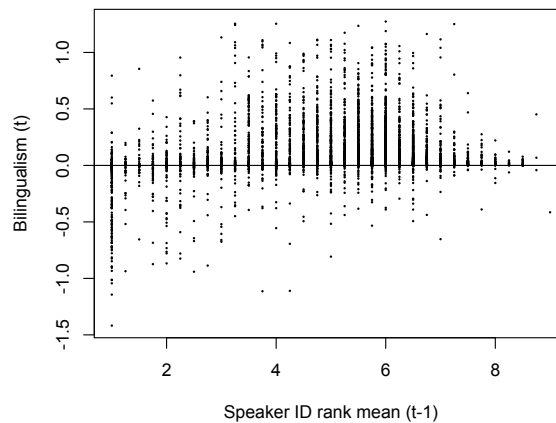


Figure C.47: The bilingualism score as a function of the mean speaker ID rank of the previous generation.

C.9.1.1 Bilingualism tracks social change

A linear regression was run to see which measures correlated best with the changes to the social structure. These measures included the bilingualism score, the rank distance (a Levenshtein distance between the ranked variables of all pairs of individuals’ linguistic models is calculated, then the mean distance between communities is taken), the vector distance (each linguistic model can be seen as a high-dimensional vector with a component for each possible variable and magnitudes being the coefficient magnitude, distances between models are the distances between these vectors) and the residual standard deviation of models (the model similarity ratio described in this section on page 240). The measures were collected for 100 independent simulations of the dynamic social structure scenario

Variable	Estimate	Std Error	t value	p	
(Intercept)	0.021	0.22	0.09	9.25E-01	
Speaker ID rank ratio (t-1)	-0.0075	0.04	-0.2	8.38E-01	
Ling. Output mean ratio (t-1)	-0.00054	0.01	-0.06	9.56E-01	
Ling Output sd ratio (t-1)	0.039	0.1	0.4	6.92E-01	
LingOutput.mean.diff (t-1)	0.086	0.05	1.73	8.36E-02	.
Model fit ratio (t-1)	-0.082	0.19	-0.43	6.68E-01	
Model similarity ratio (t-1)	0.19	0.11	1.68	9.32E-02	.
Model size ratio (t-1)	-0.21	0.12	-1.73	8.39E-02	.
Interactions ratio (t-1)	0.025	0.07	0.36	7.15E-01	
Model size mean (t-1)	-0.0097	0.01	-1.19	2.33E-01	
LingCorrBetweenGen.ratio	-0.0072	0.01	-0.56	5.76E-01	
In.tMinus1	-0.34	0.09	-3.67	2.40E-04	***
Speaker ID rank mean (t-1)	0.046	0.01	7.99	1.54E-15	***
Speaker ID rank sd (t-1)	0.064	0.03	2.51	1.22E-02	*
In.1	-0.11	0.24	-0.45	6.54E-01	
gen	0.00091	0	1.94	5.23E-02	.
Speaker rank ratio (t-1):Contact gen.	-0.019	0.04	-0.5	6.14E-01	
Ling. Output mean ratio (t-1):Contact gen.	0.00034	0.01	0.03	9.73E-01	
Ling. Output sd ratio (t-1):Contact gen.	-0.022	0.1	-0.22	8.26E-01	
Ling. Output mean diff (t-1):Contact gen.	-0.099	0.05	-1.91	5.62E-02	
Model fit ratio (t-1):Contact gen.	-0.047	0.2	-0.24	8.12E-01	
Model similarity ratio (t-1):Contact gen.	-0.18	0.12	-1.52	1.29E-01	
Model size ratio (t-1):Contact gen.	0.23	0.12	1.86	6.36E-02	.
Interactions ratio (t-1):Contact gen.	-0.016	0.07	-0.23	8.19E-01	
Model size mean (t-1):Contact gen.	0.0038	0.01	0.38	7.05E-01	
LingCorrBetweenGen.ratio:In.1	0.0077	0.01	0.53	5.98E-01	
In.tMinus1:In.1	0.48	0.1	4.68	2.90E-06	***
SpeakerRank.all.mean.tMinus1:In.1	-0.02	0.01	-2.81	4.91E-03	
SpeakerRank.all.sd.tMinus1:In.1	0.14	0.02	6.7	2.25E-11	***
In.1:gen	-0.00076	0	-1.43	1.54E-01	
Speaker ID rank (t-1) mean:sd	-0.022	0	-7.08	1.59E-12	***

Table C.1: Results of a linear regression predicting the bilingualism score.

described in section 7.7 (number of semantic variables = 4; number of peaks in the semantic distributions= 2; number of data points per individual = 100; error parameter = 1; Information criterion $k = 2$). The bilingualism score predicts the integration parameter better than some other measures of linguistic variation: In a linear model containing all variables it has the largest t value (model explained 14% of the variation $F(7543,4)=308.9$, $p < 0.0000001$).

Variable	Estimate	Std. Error	t value	p	
(Intercept)	9.010e-01	6.834e-03	131.833	<2e-16	***
Bilingualism score	4.061e-01	1.248e-02	32.541	<2e-16	***
Rank distance	-2.990e-01	1.675e-02	-17.857	<2e-16	***
Vector distance	1.562e-05	2.083e-06	7.501	7.07e-14	***
RSD	-4.898e-03	3.301e-03	-1.484	0.138	

Table C.2: Results of a linear model predicting the integration parameter I based on different measures of linguistic variation.

Also, in a linear regression, it significantly improves the fit of the model after all other variables have been entered, ($F(7543,1) = 1058$, $p < 0.000001$). This suggests that the bilingualism score reflects more than just typological distance.

C.9.2 Complex interactions

In order to discover some more complex interactions between variables, I ran a stepwise regression on the data from many simulations. The initial model included all pairwise interactions between the variables discussed above, plus those interactions with the integration parameter and generation number. The stepwise regression was run with the Bayesian Information Criterion ($\log(N)$ where $N=100$) to reduce the number of parameters in the final model. The final model accounted for 21.4% of the variation in the bilingualism score ($F(55,7343) = 37.62$, p-value: < 0.000000001). However, this still resulted in 56 parameters, 42 of which were interactions between at least two variables (see table C.3). Below are some analyses of the most significant factors. They will try to identify what differentiates generations that exhibit negative bilingualism scores from positive bilingualism scores. For convenience, the analyses are included in the caption of the graphs. The graphs include the following relationships with the bilingualism score: Speaker ID rank ratio and model size ratio; Integration and speaker ID rank means; Speaker id rank mean and standard deviation; Model fit ratio and difference in the linguistic signal.

Var. #	Variable	Estimate	Std. Error	t value	p
9	ModelSize.interaction.ratio.tMinus1	0.14	0.08	1.77	7.68e-02
39	LingOutput.sd.ratio.tMinus1:In.1	-0.34	0.16	-2.17	2.98e-02
36	In.tMinus1:SpeakerRank.all.sd.tMinus1	-0.43	0.19	-2.30	2.14e-02
4	LingOutput.sd.ratio.tMinus1	0.41	0.15	2.68	7.41e-03
8	ModelSize.ratio.tMinus1	-0.59	0.21	-2.74	6.14e-03
23	LingOutput.sd.ratio.tMinus1:SpeakerRank.all.sd.tMinus1	-0.44	0.16	-2.74	6.14e-03
38	SpeakerRank.ratio.tMinus1:In.1	-1.41	0.51	-2.75	6.00e-03
10	ModelSize.mean.tMinus1	0.07	0.02	2.79	5.36e-03
18	SpeakerRank.ratio.tMinus1:RSquareds.ratio.tMinus1	-1.47	0.51	-2.90	3.74e-03
44	ModelSize.mean.tMinus1:In.1	-0.09	0.03	-2.97	3.04e-03
49	LingOutput.sd.ratio.tMinus1:SpeakerRank.all.sd.tMinus1:In.1	0.48	0.16	2.99	2.83e-03
27	RSquareds.ratio.tMinus1:RSD.ratio.tMinus1	0.08	0.03	3.01	2.66e-03
17	SpeakerRank.ratio.tMinus1:LingOutput.mean.diff.tMinus1	-0.08	0.03	-3.01	2.62e-03
31	ModelSize.ratio.tMinus1:SpeakerRank.all.mean.tMinus1	0.09	0.03	3.09	1.99e-03
2	SpeakerRank.ratio.tMinus1	1.58	0.51	3.11	1.89e-03
48	SpeakerRank.ratio.tMinus1:RSquareds.ratio.tMinus1:In.1	1.61	0.51	3.16	1.61e-03
29	RSD.ratio.tMinus1:ModelSize.interaction.ratio.tMinus1	-0.08	0.03	-3.17	1.51e-03
7	RSD.ratio.tMinus1	0.04	0.01	3.26	1.13e-03
30	ModelSize.ratio.tMinus1:ModelSize.interaction.ratio.tMinus1	0.03	0.01	3.28	1.05e-03
13	SpeakerRank.all.sd.tMinus1	0.63	0.19	3.34	8.52e-04
25	LingOutput.mean.diff.tMinus1:ModelSize.mean.tMinus1	0.23	0.07	3.42	6.21e-04
43	ModelSize.interaction.ratio.tMinus1:In.1	-0.33	0.10	-3.46	5.36e-04
53	ModelSize.ratio.tMinus1:SpeakerRank.all.mean.tMinus1:In.1	-0.12	0.03	-3.50	4.61e-04
56	In.tMinus1:SpeakerRank.all.sd.tMinus1:In.1	0.70	0.20	3.56	3.75e-04
40	LingOutput.mean.diff.tMinus1:In.1	-5.53	1.49	-3.72	2.00e-04
15	SpeakerRank.ratio.tMinus1:LingOutput.mean.ratio.tMinus1	0.04	0.01	3.76	1.72e-04
3	LingOutput.mean.ratio.tMinus1	-0.04	0.01	-3.78	1.58e-04
28	RSquareds.ratio.tMinus1:ModelSize.interaction.ratio.tMinus1	-0.10	0.03	-3.84	1.23e-04
42	ModelSize.ratio.tMinus1:In.1	0.92	0.24	3.87	1.08e-04
51	LingOutput.mean.diff.tMinus1:ModelSize.mean.tMinus1:In.1	-0.28	0.07	-3.91	9.49e-05
6	RSquareds.ratio.tMinus1	2.30	0.58	3.95	8.00e-05
5	LingOutput.mean.diff.tMinus1	5.87	1.48	3.97	7.28e-05
41	RSquareds.ratio.tMinus1:In.1	-2.51	0.59	-4.27	1.95e-05
54	ModelSize.mean.tMinus1:SpeakerRank.all.mean.tMinus1:In.1	0.03	0.01	4.42	9.95e-06
47	SpeakerRank.all.sd.tMinus1:In.1	-0.90	0.20	-4.50	6.80e-06
32	ModelSize.interaction.ratio.tMinus1:In.tMinus1	0.30	0.06	4.60	4.37e-06
52	LingOutput.mean.diff.tMinus1:SpeakerRank.all.mean.tMinus1:In.1	-0.18	0.04	-4.69	2.82e-06
34	ModelSize.mean.tMinus1:SpeakerRank.all.sd.tMinus1	0.02	0.00	4.75	2.08e-06
26	LingOutput.mean.diff.tMinus1:SpeakerRank.all.mean.tMinus1	0.17	0.03	4.94	7.92e-07
33	ModelSize.mean.tMinus1:SpeakerRank.all.mean.tMinus1	-0.02	0.00	-4.95	7.57e-07
21	SpeakerRank.ratio.tMinus1:In.tMinus1	-0.27	0.05	-5.09	3.66e-07
46	SpeakerRank.all.mean.tMinus1:In.1	-0.34	0.07	-5.10	3.54e-07
11	In.tMinus1	2.13	0.38	5.63	1.88e-08
37	SpeakerRank.all.mean.tMinus1:SpeakerRank.all.sd.tMinus1	-0.02	0.00	-5.66	1.56e-08
22	SpeakerRank.ratio.tMinus1:SpeakerRank.all.sd.tMinus1	0.05	0.01	5.93	3.16e-09
12	SpeakerRank.all.mean.tMinus1	0.32	0.05	5.93	3.08e-09
24	LingOutput.mean.diff.tMinus1:RSquareds.ratio.tMinus1	-8.69	1.44	-6.02	1.84e-09
50	LingOutput.mean.diff.tMinus1:RSquareds.ratio.tMinus1:In.1	8.77	1.45	6.06	1.47e-09
1	(Intercept)	-4.09	0.67	-6.07	1.33e-09
14	In.1	4.36	0.70	6.24	4.76e-10
20	SpeakerRank.ratio.tMinus1:ModelSize.interaction.ratio.tMinus1	0.04	0.01	6.44	1.29e-10
45	In.tMinus1:In.1	-2.47	0.38	-6.44	1.28e-10
35	In.tMinus1:SpeakerRank.all.mean.tMinus1	-0.39	0.06	-6.56	5.89e-11
16	SpeakerRank.ratio.tMinus1:LingOutput.sd.ratio.tMinus1	-0.07	0.01	-6.88	6.62e-12
55	In.tMinus1:SpeakerRank.all.mean.tMinus1:In.1	0.45	0.06	6.97	3.45e-12
19	SpeakerRank.ratio.tMinus1:ModelSize.ratio.tMinus1	-0.15	0.02	-7.14	1.05e-12

Table C.3: Results of a stepwise regression of factors that influence the bilingualism score in a simulation with dynamic social structures. Parameters are ordered from least significant to most significant.

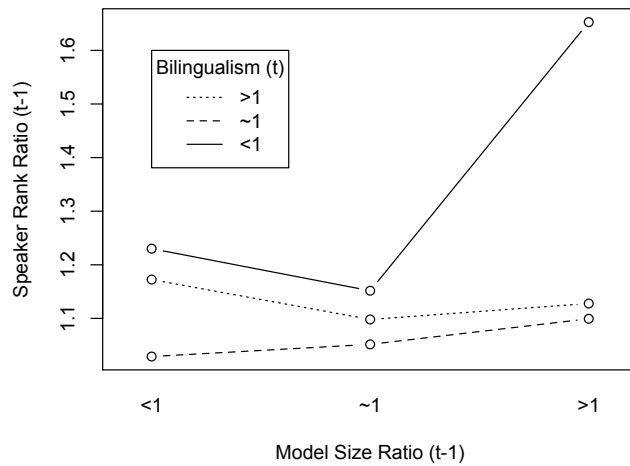


Figure C.48: Interaction between bilingualism score, speaker ID rank ratio and model size ratio. In generations preceding populations with negative bilingualism scores, speaker ID rank ratios tend to be greater than one when the model size ratios are greater than one. That is, negative bilingualism scores resulted when the community with the largest model also considered speaker ID to be less important.

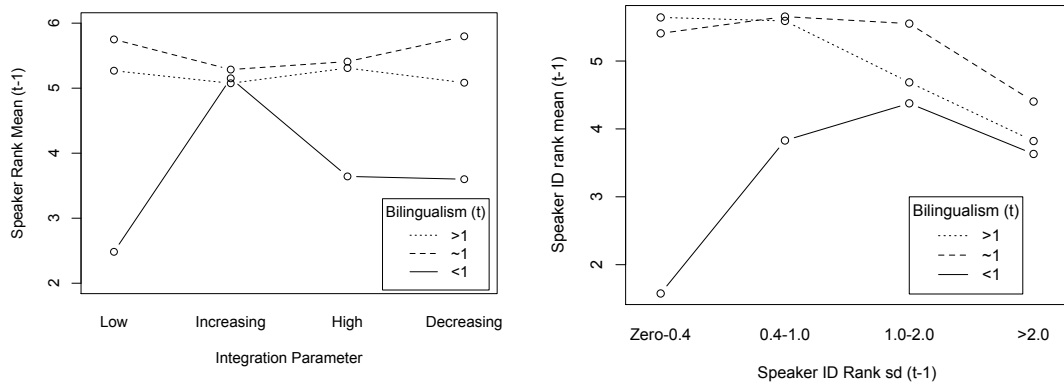


Figure C.49: **Left:** Interaction between bilingualism score, speaker ID rank mean and the change in the integration parameter. The rank of the speaker ID tends to be similar for all bilingualism types when the integration parameter is increasing (communities becoming more isolated, see figure C.49). However, in populations with a negative bilingualism score, the speaker ID rank of the previous generation tends to be lower (more important) when the integration parameter is low (fully integrated), high (isolated) or decreasing (becoming less isolated). **Right:** Interaction between bilingualism score and the mean and standard deviation of the speaker ID rank in the previous generation. Populations with negative bilingualism scores tend to be preceded by generations where the mean and standard deviations of the speaker id rank are correlated. That is, when there is little difference between the speaker id ranks of the individuals, and the speaker id rank is low (more important), then the next generation tends to have a negative bilingualism score.

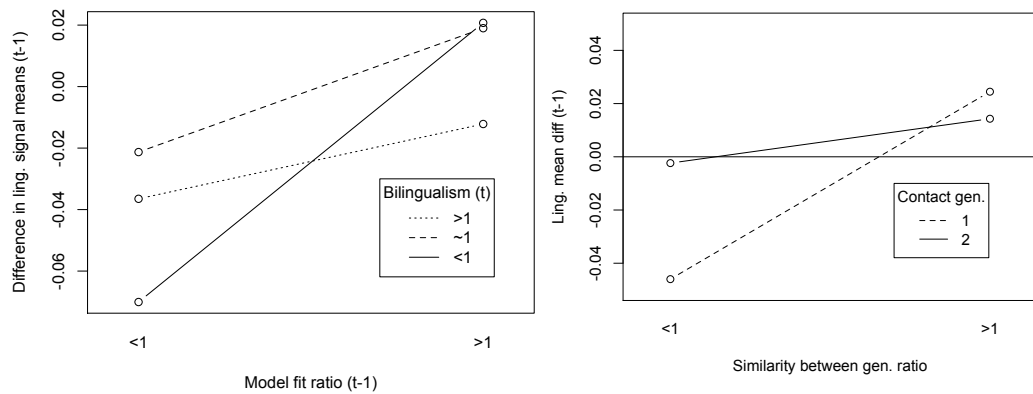


Figure C.50: **Left:** The interaction between the bilingualism score, the model fit ratio and the difference in linguistic signal means. There is a stronger correlation between the difference in linguistic signal mean and model fit ratio for populations with a negative bilingualism score than any other. That is, populations with negative bilingualism scores tend to be preceded by generations where the community with the best model fit is also the community whose linguistic signal mean has changed the least. **Right:** The interaction between the model similarity ratio between generations and the difference in linguistic means between generations, by contact generation (1=first generation of contact, 2=second generation of contact). If community A's mean is higher than community B's mean in the previous generation, then community A's language will change more than community B. However, this trend is only strong in the first generation of contact.

C.9.3 Predicting linguistic change

A linear regression was run with the linguistic correlation between generations as the dependent variable and a range of measures described above as independent variables. The regression was run separately for generations in which the communities were diverging and converging in their interaction. Table C.4 shows the results for when a population where the interaction between individuals is becoming more structured. The regression was based on 374 runs of the model and accounted for 11.2% of the variance ($F(26,373) = 1.8$, $p = 0.01$). Table C.5 shows the results for when the population is becoming more integrated. The regression was based on 176 runs of the model and accounted for 20.7% of the variance ($F(24, 175)=3.17$, $p = 0.00006$). See section 7.7.1 in chapter 7 for a discussion of the relevant results.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.02	0.6	3.39	7.81E-04	***
Speaker ID rank ratio (t-1)	-0.04	0.05	-0.85	3.98E-01	
Ling. Output mean ratio (t-1)	0.01	0.02	0.29	7.70E-01	
Ling Output sd ratio (t-1)	0.03	0.14	0.24	8.13E-01	
LingOutput.mean.diff (t-1)	-0.02	0.06	-0.35	7.27E-01	
Model fit ratio (t-1)	-1.01	0.59	-1.73	8.49E-02	.
Model similarity ratio (t-1)	-0.27	0.16	-1.69	9.14E-02	.
Model size ratio (t-1)	0.62	0.15	4.23	2.91E-05	***
Interactions ratio (t-1)	-0.36	0.09	-3.92	1.07E-04	***
Model size mean (t-1)	0	0.01	0.31	7.56E-01	
Contact gen.	-0.76	0.35	-2.2	2.85E-02	*
Speaker ID rank mean (t-1)	-0.01	0.01	-0.7	4.87E-01	
Speaker ID rank sd (t-1)	0.02	0.03	0.59	5.52E-01	
Gen.	0.0004	0.0006	0.87	3.84E-01	
Speaker rank ratio (t-1):Contact gen.	0.02	0.03	0.72	4.72E-01	
Ling. Output mean ratio (t-1):Contact gen.	-0.004	0.02	-0.25	8.00E-01	
Ling. Output sd ratio (t-1):Contact gen.	0.03	0.09	0.37	7.09E-01	
Ling. Output mean diff (t-1):Contact gen.	0.002	0.04	0.05	9.58E-01	
Model fit ratio (t-1):Contact gen.	0.67	0.34	1.97	5.01E-02	.
Model similarity ratio (t-1):Contact gen.	0.23	0.1	2.35	1.91E-02	*
Model size ratio (t-1):Contact gen.	-0.37	0.1	-3.82	1.55E-04	***
Interactions ratio (t-1):Contact gen.	0.21	0.06	3.27	1.16E-03	**
Model size mean (t-1):Contact gen.	-0.004	0.01	-0.72	4.74E-01	
Contact gen.:Speaker ID rank mean (t-1)	0.004	0	0.92	3.56E-01	
Contact gen.:Speaker ID rank sd (t-1)	-0.01	0.01	-0.45	6.52E-01	
Contact gen.:Gen.	-0.0002	0	-0.64	5.22E-01	
Speaker ID rank (t-1) mean:sd (t-1)	-0.002	0	-0.53	5.97E-01	

Table C.4: Results of a linear regression predicting the model similarity ratio for generations where the population is becoming more structured.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.42	1.75	1.38	1.68E-01	
Speaker ID rank ratio (t-1)	-0.05	0.1	-0.52	6.06E-01	
Ling. Output mean ratio (t-1)	-0.05	0.07	-0.65	5.19E-01	
Ling Output sd ratio (t-1)	-0.46	0.3	-1.54	1.26E-01	
LingOutput.mean.diff (t-1)	0.58	0.17	3.49	6.01E-04	***
Model fit ratio (t-1)	-0.42	1.61	-0.26	7.96E-01	
Model similarity ratio (t-1)	0.07	0.57	0.12	9.06E-01	
Model size ratio (t-1)	-0.33	1.09	-0.31	7.60E-01	
Interactions ratio (t-1)	0.24	0.63	0.37	7.09E-01	
Model size mean (t-1)	-0.05	0.05	-0.91	3.62E-01	
Contact gen.	-0.66	1.17	-0.57	5.72E-01	
Speaker ID rank mean (t-1)	0.01	0.03	0.26	7.93E-01	
Speaker ID rank sd (t-1)	0.02	0.14	0.12	9.08E-01	
Speaker rank ratio (t-1):Contact gen.	0.03	0.08	0.35	7.27E-01	
Ling. Output mean ratio (t-1):Contact gen.	0.04	0.06	0.7	4.85E-01	
Ling. Output sd ratio (t-1):Contact gen.	0.64	0.23	2.83	5.13E-03	**
Ling. Output mean diff (t-1):Contact gen.	-0.55	0.12	-4.7	5.34E-06	***
Model fit ratio (t-1):Contact gen.	-0.23	1.11	-0.21	8.33E-01	.
Model similarity ratio (t-1):Contact gen.	-0.18	0.36	-0.48	6.28E-01	
Model size ratio (t-1):Contact gen.	0.11	0.7	0.15	8.79E-01	
Interactions ratio (t-1):Contact gen.	-0.09	0.39	-0.23	8.17E-01	
Model size mean (t-1):Contact gen.	0.04	0.03	1.24	2.15E-01	
Contact gen.:Speaker ID rank mean (t-1)	-0.01	0.02	-0.29	7.73E-01	
Contact gen.:Speaker ID rank sd (t-1)	0.03	0.06	0.56	5.77E-01	
Speaker ID rank (t-1) mean:sd (t-1)	-0.01	0.02	-0.44	6.60E-01	

Table C.5: Results of a linear regression predicting the model similarity ratio for generations where two previously isolated populations are integrating.

APPENDIX D

MANUSCRIPTS

D.1 Monolingual Biases in Simulations of Cultural Transmission

Roberts, S. (in press). Monolingual biases in simulations of cultural transmission. In Virginia Dignum, Frank Dignum, Jacques Ferber, Tiberiu Stratulat (eds.) *Integrating Cultures*. Springer.

Monolingual Biases in Simulations of Cultural Transmission

Seán Roberts

Abstract Recent research suggests that the evolution of language is affected by the inductive biases of its learners. I suggest that there is an implicit assumption that one of these biases is to expect a single linguistic system in the input. Given the prevalence of bilingual cultures, this may not be a valid abstraction. This is illustrated by demonstrating that the ‘minimal naming game’ model, in which a shared lexicon evolves in a population of agents, includes an implicit mutual exclusivity bias. Since recent research suggests that children raised in bilingual cultures do not exhibit mutual exclusivity, the individual learning algorithm of the agents is not as abstract as it appears to be. A modification of this model demonstrates that communicative success can be achieved without mutual exclusivity. It is concluded that complex cultural phenomena, such as bilingualism, do not necessarily result from complex individual learning mechanisms. Rather, the cultural process itself can bring about this complexity.

1 Introduction

Cultural groups are very rarely isolated. They interact for trade, politics and war. Communication is key to these interactions, and so a common language is important. The emergence of common languages has been studied using computational models. However, one aspect of cultural interaction has been left largely ignored - the ability to learn many languages at once, or bilingualism. This chapter considers the importance of incorporating bilingualism into studies of cultural evolution.

Bilingualism is by no means a rare phenomenon. Statistics on the exact prevalence of bilingualism are difficult to obtain. In the USA 18% of the population are estimated to speak two or more languages [1]. The estimate is 34% for Canada[2], 66 % in the EU [3] and 80% in China [4]. Bilinguals are a majority in about a third of countries [4]. These are likely to be conservative estimates, and with over 6,000 languages squeezed into in around 200 nations, it’s likely that contact with multiple languages is an everyday feature of most people’s lives.

Recently, industrialisation and globalisation have meant that, in the first world, the perception of the prevalence of bilingualism is artificially low - especially for native speakers of global languages such as English [5, 6, 7]. It’s no surprise, then, that when cultural processes come to be modeled, one of the first simplifying assumptions would be that people speak one language. However, the abstraction to monolingualism ignores several linguistic phenomena such as the prevalence of bilingualism in societies and the ease with which children learn more than one language [8].

This chapter will consider the validity of monolingual assumptions in models of cultural evolution. Firstly, the way in which bilingualism might affect the evolutionary dynamics of language is explored. Next, a case-study of the ‘Minimal

Seán Roberts

Language Evolution and Computation, PPLS, The University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, UK, e-mail: s.g.roberts@sms.ed.ac.uk

Naming Game' will reveal an implicit monolingual bias, namely mutual exclusivity (the assumption that each object only has one name and each name only refers to one object, see [9, 10]). Since bilinguals do not exhibit mutual exclusivity [11, 12, 13], the model is generalised to weaken this constraint. The model demonstrates that communicative success can be achieved even without mutual exclusivity, in opposition to previous research [14, 15]. The model suggests that cultural phenomena adapt to the function they are required to fulfill (e.g. [16, 17]). When seeking to model the integration of cultures a common measurement is required. However, even small differences in the way different communities interact can lead to fundamental cultural differences between them, meaning that a common metric might be very abstract.

1.1 Bilingualism and cultural evolution

The dynamics of language evolution have been extensively studied through computational modelling. The canonical language learner in these models is an agent that tries to settle on a single grammar that explains the variation in its input. This implicit monolingualism is seen as a necessary abstraction in order to get at the more fundamental dynamics of language evolution. There is a sense in the field of language evolution that bilingualism is a sociolinguistic phenomenon that is the product of the interactions of several monolingual communities who have already evolved language. Implicitly, bilingualism is seen as a secondary linguistic ability - a sort of by-product.

For instance, many models represent languages as discrete entities which compete with one another [18, 19]. Even when language is modelled as distributions over words, two standard simplifying assumptions are made by many approaches to language evolution and change (e.g. [20, 21, 22]). Firstly, it is assumed that there are discrete generations with one agent per generation. This limits the amount of complexity that can be added by the cultural system. Secondly, it is assumed that all learners use the same learning algorithm, or that learning algorithms do not change over a learner's lifetime.

The first assumption has already been criticised [23, 24] and recent research has shown that the complexity of cultural dynamics can effect the eventual distribution of languages in a population [15]. A model has also been proposed which allows agents to speak and acquire multiple languages from multiple speakers [24].

However, the second assumption may also be called into question. I will illustrate this with research on the mutual exclusivity bias, and continue in the next section to show that this bias exists in certain models of language evolution and change. It has been demonstrated that monolingual children and adults exhibit a mutual exclusivity bias [9, 10]: a tendency to assume that each object only has one name and each name only refers to one object. However, recent research has shown that bilinguals do not exhibit mutual exclusivity [11, 12, 13]. It is hypothesised that the bias is overridden because of a higher variance in the input of children in bilingual contexts. Applying mutual exclusivity when presented with two languages is not suitable, since there will be at least two words for each object.

If the amount of linguistic variance (at any level of description) influences the learning strategy for that variance, then this will affect the selective pressure on languages. This will, in turn, affect the kinds of languages that emerge, thus feeding back into the amount of linguistic variance. These aspects would then co-evolve.

Given this, there are two possible fundamental states of the language learner. Either they begin with a mutual exclusivity bias which is overridden in certain situations or they begin with no assumptions and develop mutual exclusivity if the conditions are right. In the next section, it will be shown that some models make implicit assumptions about the development of mutual exclusivity and see it more as a fundamental part of language acquisition and language evolution rather than an acquired heuristic that is applied in suitable contexts. It will be argued that the most abstract learner is one without the mutual exclusivity bias, and so models should not assume mutual exclusivity as part of the learner's bias.

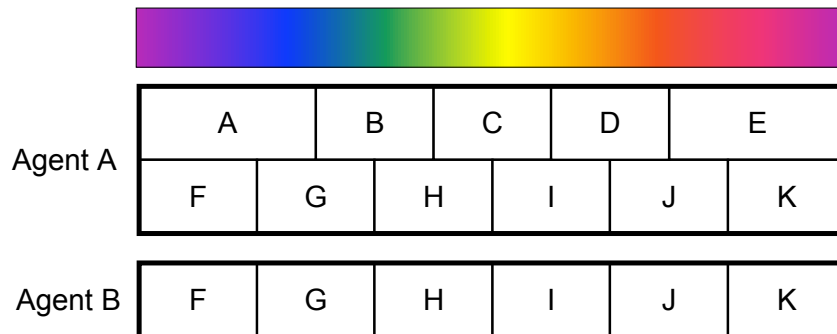


Fig. 1 An example of how two agents might split the meaning space into categories and label those categories. The meaning space spans the interval 0.0 to 1.0. Agent A and B both have the same conceptual space, but Agent A has multiple labels in each category while agent B only has one label in each category. The representation for agent A above pulls apart sections of the space that are contiguously labelled with the same label into two systems.

2 Categorisation Games

This section presents a case-study of a model of cultural evolution - the Categorisation Game - and demonstrates implicit monolingual biases that obscure some interesting dynamics. The Categorisation Game looks at how agents in a population converge on a shared system for referring to continuous stimuli [25, 26, 27]. This paradigm is often couched in terms of deciding on words for objects referred to by their colour. The colour spectrum is continuous, so agents must decide where to place category boundaries as well as the label for that category. The ‘minimal naming game’ [28] (also used in [29, 30, 31]) is a simplification of the categorisation game which “possibly represents the simplest example of the complex processes leading progressively to the establishment of complex human-like languages” [28]. I’ll show that even this ‘minimal’ algorithm has implicit monolingual assumptions. First, however, a note is made about the measurements that researchers have used to study the categorisation game.

2.1 Measurements of coherence

Other models looking at this problem have considered measurements apart from communicative success. For instance, the ‘level of lexical coherence’ in the system, according to [32] is the average proportion of shared lexical items in a population. The category overlap function [28, 30] measures the level of alignment between the category boundaries of the agents. However, an appropriate measurement when considering the possibility of ‘bilingualism’ is less clear. For instance, consider the example of two agents with categories and labels as described in figure 1. Adapting the lexical coherence measurement from [32] gives a coherence of 75%. This measurement fails to capture the fact that agent B would always be understood by agent A and that agent A could always make itself understood to agent B given the right choice of lexical item. In other words, although the agents have differences in the words that they know, they are still able to communicate unambiguously about the whole spectrum.

Measuring category overlap is also problematic. Agents with category boundaries at exactly the same locations will have a category overlap of 1.0. However, the overlap of the example above is 0.09, despite the relatively good communicative success possible between the pair. This is because the measurement collapses the category boundaries of an agent into a single system before comparing it to another agent. By doing this, the division between the two ‘languages’ of agent A in figure 1 is ignored.

These measures reflect the level of coherence in the population, but only effectively for a population whose goal it is to converge on a single, ‘monolingual’ system. Researchers have used these measurements to gauge the progress of their

model, demonstrating a monolingual bias in their approach. Further research is required to find a good way of measuring coherence in a heterogenous population (see [33, 41, 34]). This paper will proceed assuming that communicative success should be the most important measure of coherence between agents.

2.2 The minimal naming game

The algorithm for the categorisation game is reproduced below. However, two of the steps are re-analysed as heuristics rather than essential elements. These heuristics impose a mutual exclusivity bias in the agents. The steps are as follows (following [30]): There is a population of N agents, each able to partition the perceptual space into categories. Each category has a list of associated words. Each agent has a minimum perceptual difference threshold d_{min} , below which stimuli appear the same. At each time step:

1. Two individuals are chosen at random to be the speaker and the listener.
2. They both have access to a scene containing M stimuli. The stimuli must be perceptually distinguishable by the agents (perceptual distance $\geq d_{min}$).
3. The speaker selects a topic and discriminates it in the following way:
 - Each stimulus is assigned to a perceptual category
 - If one or more other stimuli are assigned to the same category as the topic, the agent splits its perceptual categories so that each stimulus belongs to only one perceptual category. Within a category with two or more stimuli, a boundary is placed halfway between the first two stimuli.
 - The new partitions inherit the associated words of the old partition.
 - **Heuristic A:** Each new partition is given a new, unique name. It's assumed that no two agents will create the same name.
4. The speaker transmits a word that it associates with the topic to the listener. If it has no words associated with the category, it creates a new one. If it has more than one word associated, it transmits the one that was last used in a successful communication.
5. The hearer receives the word and finds all categories which have the associated word and which identify one of the stimuli in the scene. Then:
 - If there are no such categories, the agent does nothing.
 - If there is one such category, the agent points to the associated stimulus.
 - If there is more than one such category, the agent points randomly at an associated stimulus.
6. The hearer discriminates the scene, as above.
7. The speaker reveals the topic to the listener.
8. If the hearer did not point to the topic, the communication is a failure. The hearer adds the transmitted word to the category discriminating the topic.
9. If the hearer pointed to the topic, the communication is a success.

Heuristic B: Both agents delete all other words but the transmitted one from the inventory of the category discriminating the topic.

Heuristic A, above, invents new words for each sub-category when a category is split. This is an implementation of the assumption that each name only refers to one object, hence when there are two objects with the same name, the agent should discriminate between them linguistically. This interacts with Heuristic B which removes all competing names associated with a category from the listener's lexicon when communication is successful. The effect is that the listener conforms to the speaker's labeling, but also 'forgets' any previously associated words. This is an implementation of the assumption that each object only has one name.

These two heuristics, then, implement a mutual exclusivity bias: Each name only refers to one object and each object is only labeled by one name. Stable bilingualism is impossible in this model because only one name is retained after successful communication. The role of the two heuristics in the evolution of a shared communicative system is clear:

heuristic A creates new labels for categories, introducing variation into the system needed to distinguish between categories. Heuristic B causes the agents to converge on shared labels for categories by selecting for labels common to an interacting pair.

However, these heuristics are still arbitrary. As we have seen, not all human learners assume mutual exclusivity. In the next section, it will be demonstrated that a population of agents can converge on a shared communication system without these heuristics.

3 Convergence without mutual exclusivity

The algorithm was modified to remove the mutual exclusivity bias in order to test the effects on communicative success. However, the changes to the dynamics will not be explored in detail. The purpose of the changes, here, is not to explore the best way of modelling the cultural evolution of language, but to demonstrate that the biases of the researcher can influence the dynamics of the model and thus the conclusions drawn from it.

Heuristic B can be modified while retaining communicative success [35]. If the hearer, but not the speaker applies heuristic B, a coherent vocabulary still emerges in a similar time with similar memory resources required. If only the speaker applies heuristic B a coherent vocabulary does emerge, but on a longer timescale and in a qualitatively different way (approached as a thermodynamic system, consensus is reached due to large, system-size fluctuations of the magnetisation [35]). However, this research was concerned with the effect of feedback on the convergence dynamics. This study looks at the assumptions built in to the individual learning algorithm.

The heuristics were modified by generalising the algorithm. Firstly, agents in a population either all applied heuristic A or all did not apply heuristic A. Heuristic B was made optional in the same way. If heuristic B did not apply, a maximum number of words s_{MAX} were retained after a successful communication. A first-in, first-out stack memory was also implemented so that the oldest stored form would be removed first. A word was pushed further back in the stack (safer from deletion) when a listener heard it being used by a speaker. This is a generalisation of the mechanism that weakens links between signals and meanings which do not co-occur.

The purpose of generalising the model was to allow bilingualism. However, the advantages of knowing more than one word for an object are not yet fully available. A bilingual, failing to communicate with one word, might try another. Therefore, the algorithm was modified to allow an arbitrary maximum number of attempts a_{MAX} at communicating before communication failed. If speakers had more than one label for a perceptual category, they transmitted them in a random order until this maximum was reached. Listeners searched their lexicon at each attempt until either they found a match in their own lexicon and made a guess at the referent or the maximum number of attempts was reached and they signaled failure, as before. Each guess was independent of any other, so successful communication was not always guaranteed, even when $a_{MAX} = M$.

It has been shown that an algorithm which leads to successful communication in a population of agents must strengthen connections between signals and meanings that appear together (or are absent together) and weaken connections between signals and meanings that do not co-occur[14]. The changes to the algorithm above do not violate these conditions, but simply weaken their strength.

4 Results

Four versions of the algorithm were run: with both heuristics, as in the original, with only heuristic A, only heuristic B and with neither heuristic. Results shown here are for a population of 4 over 10,000 rounds with a context size of 2.

Heuristics	Max	t	p	Final	t	p
A and B	0.84	23.89 ^a	< 0.0001	0.62	3.06 ^a	0.01
A only	0.95	44.26 ^b	< 0.0001	0.86	7.17 ^b	< 0.0001
B only	0.77	9.04 ^a	< 0.0001	0.52	1.44 ^a	0.18 (N.S.)
None	0.94	43.50 ^b	< 0.0001	0.84	5.96 ^b	< 0.001

^aCalculated with chance level at 0.5.

^bCalculated with chance level at 0.75.

Table 1 Communicative success for different heuristics for 10 runs each with 4 agents and $a_{MAX}=2$ and $s_{MAX}=2$ for 10,000 rounds. The maximum communicative success was calculated as follows: for each run, the average communicative success for data grouped into 100-round bins. The average of a bin was taken as the value of that bin. The values shown in the table are the average of the maximum bin values for 10 runs. The final communicative success is the average success for last 100 rounds over 10 runs. The statistics show a two tailed t-test indicating performance above chance.

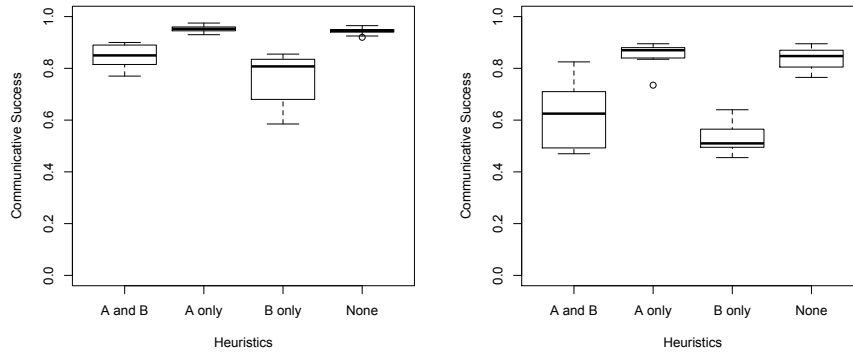


Fig. 2 Maximum communicative success (left) and final communicative success after 10,000 rounds (right) for ten runs of populations with various heuristics. Statistics in table 1.

4.1 Communicative Success

Table 1 and figure 2 show the communicative success for the algorithm run with different heuristics with $a_{MAX}=2$ and $s_{MAX}=2$. All heuristics manage in achieving good communicative success at some point (shown by the maximum communicative success achieved). That is, a mutual exclusivity bias is not necessary for communicative success in this model. It should be noted that the probability of choosing the correct referent by chance is $\frac{1}{c} = 0.5$ (where c is the context size) because the algorithm tends to limit the number of words linked to a perceptual category to one. However, algorithms without heuristic B (i.e. ‘A only’ and ‘None’) have a higher probability because they are more likely to be able to take advantage of extra communicative attempts. Therefore, for algorithms without heuristic B, the probability of selecting the correct referent by chance is

$$\sum_{i=1}^{a_{MAX}} \left(\frac{1}{c}\right)^i \quad (1)$$

For the current settings, this is 0.75. Even taking this into account, all algorithms are able to reach stable periods with high levels of communicative success. The result is robust against changes to s_{MAX} : The relative communicative success between the different heuristic combinations remains the same for s_{MAX} up to 1000, while the absolute communicative success drops about 5% for s_{MAX} of 4 and remains around that level for s_{MAX} up to 1000.

However, eventually all agents converge on a single word for the whole meaning space. This is typical behaviour for this model [36]. This reduces the communicative success, since agents cannot distinguish linguistically between referents. Table 1 shows the average final communicative success after 10,000 rounds. These are less than the maximum. In the

21	61
----	----

Agent 2 (No heuristics):

0			
1	3	5	4

Fig. 3 The linguistic labels of an agent after 3,500 rounds for separate runs. Agent 1 (above) used heuristics A and B and agent 2 (below) used neither heuristic. The perceptual space runs from left to right. Contiguous linguistic categories are indicated by boxes with the linguistic label (a number in this implementation) drawn in the centre. Agent 2 has more than one label for a given perceptual stimulus.

case of using heuristic B only, the communicative success is no better than chance. The other algorithms still yield a communicative success above chance, but the algorithms without heuristic B (A only and no heuristics) do better than algorithms with heuristic B (average with B = 0.57, without B = 0.85, $t = 10.9$, $p < 0.0001$). The same collapsing process occurs as in the algorithm without heuristic B, but since there is more variation within perceptual categories due to extra labels being stored, a single label takes longer to dominate. In fact, a single linguistic item tends to spread over the whole meaning space as with the original algorithm, but a sort of secondary ‘language’ keeps distinctions between perceptual categories for longer.

Figure 3 illustrates this with a diagram of agents’ memories from mid-way through separate runs. Agent 1 was run in a population using both heuristics and agent 2 was run in a population using neither heuristic. Agent 1’s linguistic categories are already heavily collapsed while Agent 2 has a greater variation which allows it to communicate more effectively. The memories of both agents at this point are nearly perfectly similar to the other agents that they interact with.

Another measure of communicative efficiency is the entropy efficiency of an agent. Effectively, this is the average probability that an agent has a different linguistic label for any two stimuli. An agent has a set of linguistic labels which uniquely identify regions of the meaning space. L is the list of lengths of these regions. The entropy efficiency is given as

$$-\sum_{i=1}^{|L|} \frac{L_i \log(L_i)}{\log(1/d_{min})} \quad (2)$$

Since d_{min} is set so that there can be a maximum of 10 perceptually distinct regions, the highest entropy efficiency is given by an agent who can uniquely label 10 regions of equal length (entropy efficiency of 1.0). The lowest possible entropy efficiency is given by an agent with no labels or one label spanning the whole meaning space (entropy efficiency of zero). Figure 4 shows that the algorithm with both heuristics achieves a lower entropy efficiency than the algorithm without heuristic B and degrades faster than the algorithm without heuristic A.

4.2 The development of mutual exclusivity

The model has shown that mutual exclusivity is not necessary for communicative success. However, the mutual exclusivity bias is exhibited by monolinguals. The model can be manipulated to explore the rationale behind this and the most likely starting assumptions of a language learner.

Simulations were run where the mutual exclusivity heuristics were ‘switched on’ after some rounds. Figure 5 shows the difference between an algorithm that has no heuristics and one that changes to incorporate them after 1,500 rounds. For a population of two agents (low cultural complexity), switching on the heuristics makes no difference to the commu-

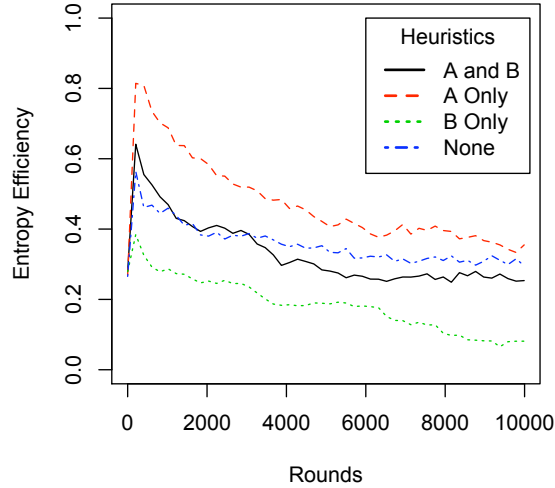


Fig. 4 Entropy efficiency for populations of agents with different heuristics. Number of agents = 4, $a_{MAX}=2$ and $s_{MAX}=2$.

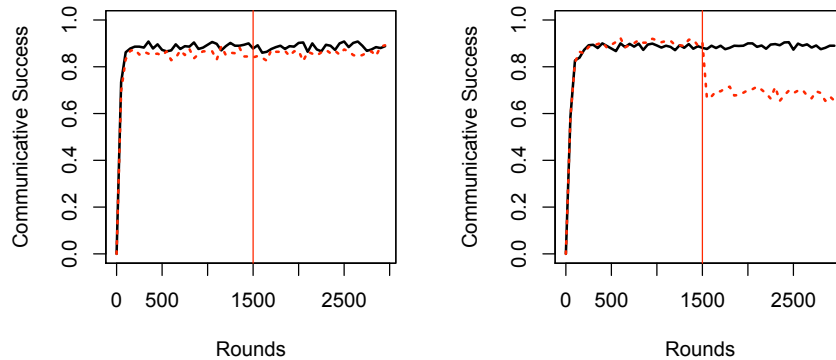


Fig. 5 Communicative success for a population of 2 agents (left) and a population of 4 agents (right). Solid lines indicate success for a consistent algorithm where no heuristics are applied. Dashed lines represent success for an algorithm that incorporates the heuristics after 1,500 rounds.

nicative success. Therefore, in this situation, applying mutual exclusivity makes rational sense in order to save memory: The application of heuristic B will reduce the number of words stored for each category. However, in a population of 4 agents, switching on the heuristics decreases the communicative success. In this situation, the most rational approach is to keep the heuristics switched 'off'. This is because the complexity of the cultural system is greater with 4 agents, leading to more variation between agents. The system evolves to store many words for an object to cope with this variation. The drop in this difference reflects the empirical findings that bilinguals do not exhibit mutual exclusivity. Figure 6 shows that this difference increases with larger populations. However, when s_{MAX} becomes many times greater than the number of agents, the disadvantage of switching decreases. That is, agents retain words that have already been discarded by others.

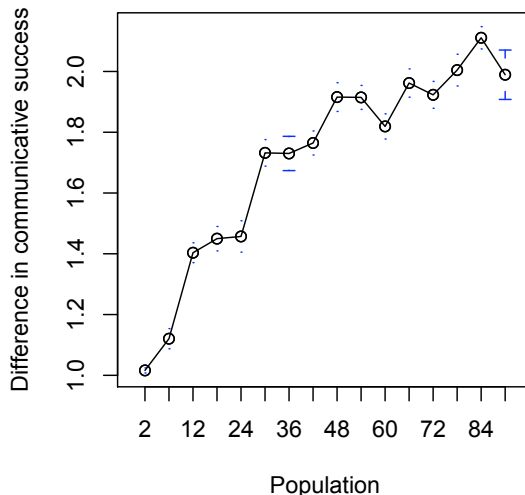


Fig. 6 The difference in percentage communicative success when switching from using no heuristics to using both heuristics for different population sizes (the difference between solid and dashed lines after 1500 rounds in figure 5). Values shown are mean and 99% confidence intervals for 5 runs. Positive values indicate an advantage for using no heuristics. In a population of 2, there is no difference between using the heuristics or not, as shown in figure 5.

The most rational strategy for any agent is not to assume mutual exclusivity to begin with, and only to activate it under relevant conditions. This reflects the findings that 14-month-old children do not exhibit it while 17-month-olds do [37]. From this model we might conclude that mutual exclusivity is an acquired heuristic which is applicable in situations where there is likely to be low variation (monolingualism). More research is required into this kind of model. The point here is that the assumptions of the original model obscure the distinction between mutual exclusivity as an innate, universal bias and an acquired, culture-specific one.

5 Discussion

Communicative success can emerge without mutual exclusivity. The results of this model stand in opposition to previous research (e.g. [14, 38, 39, 40, 41]). For instance, it has been claimed that “human language learners appear to bring a one-to-one bias to the acquisition of vocabulary systems. The functionality of human vocabulary may therefore be a consequence of the biases of human language learners” [42, p. 127]. The current research suggests that mutual exclusivity is not an innate bias. Furthermore, the bias becomes functional as a consequence of the variance in the vocabulary and social dynamic. A related model shows similar results [43]: Mutual exclusivity is not necessary for communicative success, but helps agents co-ordinate linguistically when they have conceptual differences. Multiple consensus systems can be maintained in a population with complex social structures [44]. However, the current model shows that mutual exclusivity does not always aid the co-ordination process.

However, rather than directly opposing the claims of some previous models, the constraints in the current model can be seen as a relaxation of the constraints embodied by the mutual exclusivity bias. Both models contribute the necessary ingredients for an evolutionary system: Heredity, variation and differential fitness (e.g. [45]). Although generational turnover is not modelled, there is heredity in the sense that each agent inherits its own memory from the previous round. Heuristic A introduces the variation by adding new words. Heuristic B introduces differential fitness by selecting words

which are successful in communication. Without heuristic A, variation is still introduced by agents creating new words at early stages of the game when they have no words at all (step 4 of the algorithm). The generalisation of Heuristic B to keep an arbitrary number of words after successful communication allows selection to operate over *groups* of words rather than single ones.

Heuristics A and B, then, are an efficient way of introducing the ingredients for evolution into the system. However, cultural processes can also introduce these ingredients - the individual learning processes need not be the source. Other processes could also introduce variation such as errors in production or perception or differences in contact with other agents.

6 Conclusion

The naming game was reanalysed in the light of evidence from bilingual language acquisition research. The measurements used to analyse the model were also re-assessed and shown to favour monolingual systems. Steps in the categorisation game were re-analysed as implementing a mutual exclusivity constraint. To explore the effects of these steps, the learning algorithm was generalised so that the steps could be omitted. Communicative success at the lexical level was achieved without mutual exclusivity constraints. In fact, in some cases, the constraint impedes the process.

This goes against some previous research which argued that mutual exclusivity is necessary for communication to emerge. What seems to be important is the presence of the ingredients for evolution - inheritance, variation and selection. The mutual exclusivity bias is seen as an efficient way of integrating these ingredients. However, the model also showed that rational agents should not assume mutual exclusivity to begin with. This reflects research which shows that children only start using mutual exclusivity in certain situations. Mutual exclusivity is not appropriate in a bilingual environment, so bilinguals do not exhibit it. Given this, the monolingual assumptions of the naming game are unrealistic for two reasons. First, a learner's learning algorithm may change over time, as demonstrated by the differences found between monolinguals and bilinguals. Secondly, they are not valid abstractions because the heuristics which implement mutual exclusivity are optional extras, so the simplest, default assumptions of learners should be bilingual. That is, monolingualism is a specialised form of bilingualism.

When modelling cultural processes, abstraction is necessary. However, the cultural phenomena that appear simplest (e.g. monolingualism) may not be caused by the simplest learning mechanisms. Much of the complexity in cultural phenomena stem from complex interactions *between* individuals. That is, the cultural transmission process itself can shape and influence the cultural practices it transmits.

6.1 Integrating cultures in the light of cultural adaptation

The communication system in the model above adapts to fit the needs and constraints of its users. Indeed, the hallmark of a cultural phenomenon is that it has adapted to the cognitive niche of its community's members [16, 17]. If different communities have different dynamics, such as population size or differences in social structures, then the cultural phenomena that emerge in them may be radically different. In the model above, the communication system between two agents became optimised for efficiency while the communication system in a more complex social structure became optimised for flexibility. Biases in communities towards these different optimisations could be amplified by cultural transmission [46]. Over many generations, and for a more complex cultural phenomenon (e.g. a language system, judicial system or musical form), the commonalities between two communities may erode to very abstract principles. When seeking to integrate them, then, a common measure for separate cultures may be difficult to find. Even something as simple as assuming each object only has one associated word may reflect the deep structure of the culture in which it is embedded.

Acknowledgements Thank you to Simon Kirby, Kenny Smith, Antonella Sorace and Liz Irvine for comments. Supported by the Economic and Social Research Council [ES/G010277/1]

References

1. U.S. Census Bureau: Language Use and English-Speaking Ability: 2000 (C2KBR-29). Washington: Government Printing Office. <http://www.census.gov/prod/2003pubs/c2kbr-29.pdf> (3/11/2010). (2003)
2. Statistics Canada: The Evolving Linguistic Portrait, 2006 Census Statistics Canada Catalogue no. 97-555-XIE. Ottawa. Version updated May 2010. <http://www12.statcan.ca/census-recensement/2006/as-sa/97-555/pdf/97-555-XIE2006001-eng.pdf> (3/11/2010). (2007)
3. European Commission. 2006. Special Eurobarometer: Europeans and their Languages: European Commission. http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf. cited 3 Nov 2010.
4. Baker, C. & Jones, S. P.: Encyclopedia of bilingualism and bilingual education. Clevedon, UK: Multilingual Matters. (1998)
5. Thomas, L., & Wareing, S.: Language, Society and Power. London: Routledge. (1999)
6. Kostoulas-Makrakis, N.: Language and Society. Basic Concepts (in Greek). Athens: Metexmio. (2001)
7. Luchtenberg, S.: Bilingualism and bilingual education and their relationship to citizenship from a comparative German-Australian perspective, *Intercultural Education*, Vol. 13(1), 49-61. (2002)
8. Pearson, B.Z., Fernandez, S.C., Oller, D.K.: Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning*, 43, 93-120. (1993)
9. Markman, E. M. and Wachtel, G. A.: Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121-57. (1988)
10. Merriman, W. E. and Bowman, L. L.: The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, 54 (3-4). (1989)
11. Byers-Heinlein, K. & Werker, J. F.: Monolingual, bilingual, trilingual: infants language experience influences the development of a word-learning heuristic. *Developmental Science*, 12(9), 815-823. (2009)
12. Healey, E. & Skarabela, B.: Are children willing to accept two labels for one object?: A comparative study of mutual exclusivity in bilingual and monolingual children. In *Proceedings of the Child Language Seminar*. University of Reading. (2009)
13. Houston-Price, C., Caloghiris, Z., & Raviglione, E.: Language Experience Shapes the Development of the Mutual Exclusivity Bias. *Infancy* 15 (2), 125-150. (2010)
14. Smith, K.: The cultural evolution of communication in a population of neural networks. *Connection Science*, 14, 65-84. (2002)
15. Smith, K.: Iterated learning in populations of Bayesian agents. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. (2009)
16. Christiansen, M. H. & Chater, N.: Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 5, 489-508; discussion 509-58. (2008)
17. The "Five Graces Group": Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. Language Is a Complex Adaptive System: Position Paper. *Language Learning*, 59, 1-26 (2009)
18. Niyogi, P. & Berwick, R. C.: The Logical Problem of Language Change. CBCL Paper 115, MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences. (1995)
19. Abrams, D.M. and Strogatz, S. H.: Modelling the dynamics of language death. *Nature*, 424, pp. 900. (2003)
20. Griffiths, T. L. & Kalish, M. L.: Language Evolution by Iterated Learning With Bayesian Agents. *Cognitive Science: A Multidisciplinary Journal*, 31 (3) 441-480. (2007)
21. Kirby, S.: Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5 (2) 102-110. (2001)
22. Smith, K., Kirby, S. & Brighton, H.: Iterated learning: a framework for the emergence of language. *Artificial Life* 9 (4) 371-86. (2003)
23. Niyogi, P. and Berwick, R.C.: The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106 (25) 10124-9. (2009)
24. Burkett, D., Griffiths, T.: Iterated learning of multiple languages from multiple teachers. In Smith, A., Schouwstra, M., de Boer, B., and Smith, K., editors, *The Evolution of Language: Proceedings of EvoLang 2010*, 58-65. (2010)
25. Nowak, M. A., & Krakauer, D. C.: The evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 96 (14) 8028-8033. (1999)
26. Steels, L.: Perceptually grounded meaning creation in M. Tokoro (Ed.) *Proceedings of the International Conference on Multi-agent Systems*. MIT Press. Cambridge, MA. (1996)
27. Steels, L. and Belpaeme, T.: Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28(4):469-489. (2005)
28. Loreto, V., Baronchelli, A., and Puglisi, A.: Mathematical Modeling of Language Games, Evolution of Communication and Language in Embodied Agents, part 3, 263-281. Springer Berlin Heidelberg. (2010)
29. Gong, T., Puglisi, A., Loreto, V., & Wang, W. S-Y.: Conventionalization of linguistic knowledge under simple communicative constraints. *Biological Theory: Integrating Development, Evolution, and Cognition*, 3 (2) 154-163. (2008)
30. Puglisi, A., Baronchelli, A. & Loreto, V.: Cultural route to the emergence of linguistic categories *Proceedings of the National Academy of Science*, 105 7936-7940. (2008)

31. Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V.: Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107(6):2403-2407. (2010)
32. Baronchelli, A., Felici, M. Loreto, V., Caglioti, E. & Luc Steels: Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*. P06014. (2006)
33. Komarova, N. L. & Jameson, K.A. Population heterogeneity and color stimulus heterogeneity in agent-based color categorization. *Journal of Theoretical Biology*. 253(4): 680-700. (2008)
34. De Beule, J. and De Vylder, B. and Belpaeme, T. A cross-situational learning algorithm for damping homonymy in the guessing game. In L. M. Rocha, L. Yaeger, M. Bedau, D. Floreano, R. Goldstone, and A. Vespignani (eds.) *Artificial Life X*, 466-472. MIT Press. (2006)
35. Baronchelli, A., : Do you agree? Role of feedback and broadcasting in the Naming Game. (preprint arXiv:1009.4798). (submitted)
36. Baronchelli, A., Dall'Asta, L., Barrat, A. & Loreto, V.: Bootstrapping communication in language games in A. Cangelosi, A. D. M. Smith & K. Smith, eds.: *The Evolution of Language*, Proceedings of the 6th International Conference (EVOLANG6), World Scientific Publishing Company. (2006)
37. Halberda, J.: The development of a word-learning strategy. *Cognition*, 87 (1) 23-34. (2003)
38. Vogt, P. & Haasdijk, E.: Modeling social learning of language and skills. *Artificial Life* 16 (4) 289-309. (2010)
39. Hutchins, E., & Hazlehurst, B.: How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (eds.) *Artificial societies: The computer simulation of social life*, pp. 157-189, London: UCL Press. (1995)
40. Oliphant, M.: The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, 7, 371-384. (1999)
41. De Vylder, B. & Tuyls, K.: How to reach linguistic consensus: A proof of convergence for the naming game. *Journal of Theoretical Biology*, 242 (4) 818 - 831. (2006)
42. Smith, K.: The evolution of vocabulary. *Journal of Theoretical Biology*. 228 (1) 127 - 142. (2004)
43. Smith, A. D. M.: Mutual exclusivity: communicative success despite conceptual divergence. In M. Tallerman (ed.) *Language Origins: perspectives on evolution*. 372-388. Oxford University Press. (2005)
44. de Vylder, B. *The Evolution of Conventions in Multi-Agent Systems*. PhD thesis, Artificial Intelligence Laboratory, Vrije Universiteit Brussels. (2007)
45. Lewontin, R. C.: The Units of Selection. *Annual Reviews of Ecology and Systematics* 1: 118. (1970)
46. Kirby, S., Dowman, M. & Griffiths, T.: Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104:12, 5241-5245. (2007)

D.2 Mutual exclusivity in the emergence of a shared lexicon: A pilot laboratory experiment

Roberts, S. (2011). Mutual exclusivity in the emergence of a shared lexicon: A pilot laboratory experiment. Technical report, University of Edinburgh.

Mutual exclusivity in the emergence of a shared lexicon: A pilot laboratory experiment

Seán Roberts

1 Introduction

The Naming Game is a game where players must communicate about stimuli in a continuous meaning space. This is often couched in terms of names for coloured objects. Through a process of trial and error, words come to be associated with a range of the spectrum, and these associations are shared between players.

I have shown that computational agents in the naming game do not need to apply mutual exclusivity in order to converge on a successful communication system. Furthermore, in a complex network, a mutual exclusivity constraint may hinder communicative success. The difference may be due to the optimal approaches to simple and complex networks. In a simple network with two people one has complete knowledge of the interactions of the other agent. An optimal approach is to work to find a single shared lexicon. However, in a more complex network, one does not have access to the interaction history. A good approach here may be to be more accommodating and retain several lexicons for use with each individual. That is, the speaker's identity becomes important.

So far, research on the Naming Game has been done with agent-based modelling. This experiment presents the same task to human participants. The primary question is whether participants behave differently in different networks. Specifically, whether different kinds of network encourage the emergence of mutual exclusivity behaviour.

Since this is the first human Naming Game that I am aware of, it will also be important to compare the behaviour of humans and the computational agents.

2 Estimating human learning algorithms with confederate computational agents

This experiment would like to demonstrate that human participants are more likely to develop a mutual exclusivity bias in a simple social network (2 players) than in a complex social network (4 players). However, because the learning algorithms of humans are not transparent, a method of estimating them is required. The current study is concerned with whether humans are applying a mutually exclusivity bias while playing the naming game. One way to test this is by having a number of human participants play with a computational confederate whose learning algorithm *is* transparent. By manipulating the confederate, differences in the communicative success of the group might reveal the learning algorithm of the humans.

In order to do this, two assumptions are made. Firstly, the humans in the same population will have the same learning algorithm. Secondly, that the humans will respond to a computational agent in the same way as a human agent.

Given these assumptions, adding either an agent with or without a mutual exclusivity bias will create a heterogeneous group. The computational model was run (with only computational agents) to compare performance in homogenous and heterogenous groups. In populations of two and four, a homogenous group always had a higher communicative success than a heterogenous group (see figures 1 and 2). There was no significant difference in the communicative success of agents with majority and minority algorithms.

Therefore, three conditions for each population size (2 player and 4 player) would be run: A baseline condition where the entire population was human and two test conditions which would involve one computational confederate. In the first test condition the computational agent would have a mutual exclusivity bias and the second test condition the computational agent would have no mutual exclusivity bias. It is assumed that the test condition in which communicative success was better would be a heterogenous group. That is, the group that does better has the same learning algorithm as its computational confederate.

It is predicted that, in the two player condition, the best performance comes from interacting with an agent with a mutual exclusivity bias. In the four player condition, the best performance is predicted to come from interacting with an agent without a mutual exclusivity bias.

One possible confound is the ratio of human to computational players. In the two player condition, half of the population is human while in the four player condition two thirds are human.

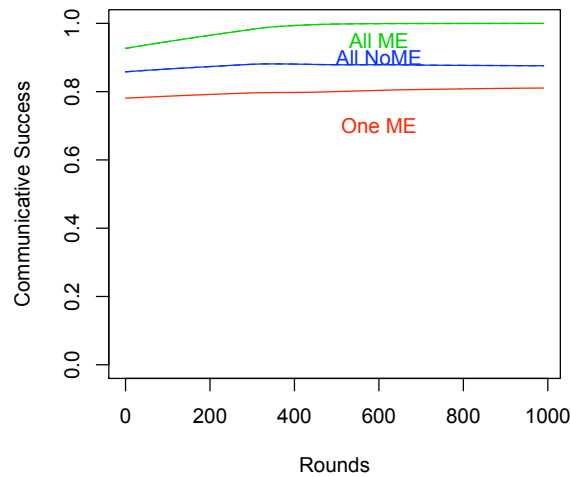


Figure 1: Communicative success for agents in homogenous and heterogenous populations with 2 agents.

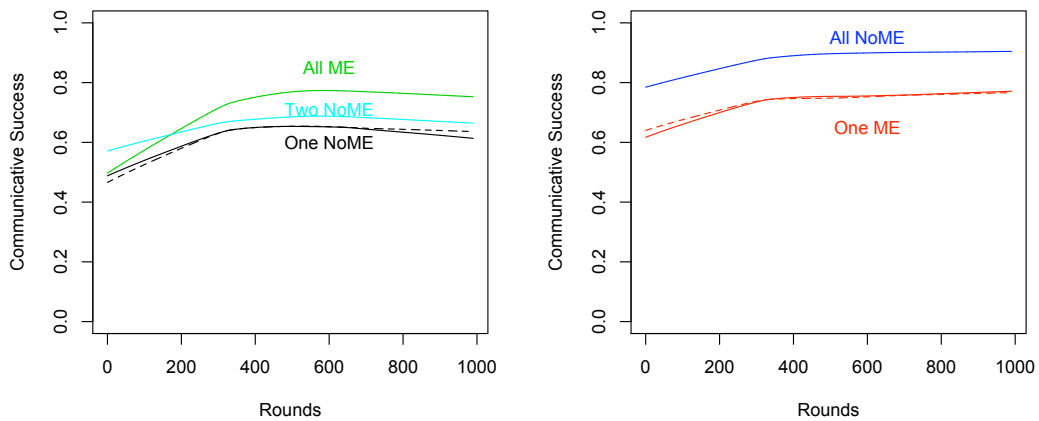


Figure 2: Communicative success for agents in homogenous and heterogenous populations with 4 agents. The dashed line shows the communicative success of the agent with the minority algorithm. The data is split into two graphs for clarity.

3 The Naming Game with Human Participants

This experiment replicates the Naming Game with human participants. Participants are presented with a context of several coloured squares and a fixed set of words that they can use to communicate about them. The main dependent variable of the experiment is social network type. In other words, how many other people you communicate with. A simple network contains two players who communicate with each other all the time. A complex network contains four players, each of whom interacts with each other player a third of the time.

4 Predictions for the lexicon

For the experiment with only human players, it is predicted that the approach to learning will be affected by the complexity of the social network. However, the eventual communicative success will be equally high. Lexicons will develop differently in the simple and complex networks. Specifically, lexicons in the complex network will be larger than in the simple network. Participants in the complex network may remember more than one label for a given range of colours to accommodate separate interaction histories.

Participants in the simple network are likely to develop mutual exclusivity behaviour. However, this is less appropriate in the complex network. Therefore, it is predicted that participants in the complex network will be more willing to violate mutual exclusivity. This will be tested using a mutual exclusivity test (similar to the DAX test) at the end of the experiment.

5 Method

5.1 Stimuli

5.1.1 Colours

A key parameter in the model is d_{min} - the distance at which two colours become perceptually indistinguishable. The colours presented to the human participants in the context must be at least perceptually distinguishable, but human sensitivity to colour is not uniform over the spectrum. Figure 4 shows the difference in wavelength needed to cause a Just Noticeable Difference (JND) across the spectrum (from Long et al., 2006). The human eye is better at discriminating colours in the cyan and yellow areas than the red and purple areas. Data on the average human JND from Long et al. (2006) was used to define a colour space with approximately uniform perceptual difference.

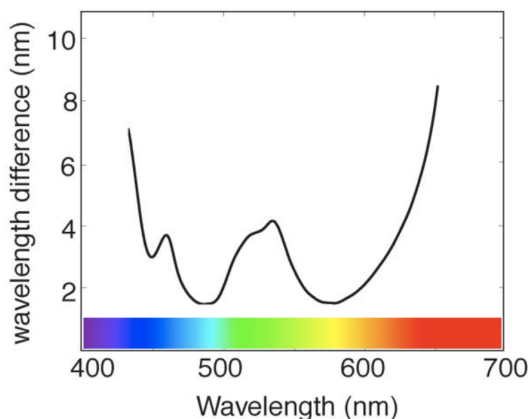


Figure 3: The wavelength change in a monochromatic stimulus needed to elicit a just noticeable difference in hue (Long et al., 2006).



Figure 4: Twenty evenly spaced colours according to the human JND (top) and a physical (linear) colour scale (bottom). The transitions between each colour should appear smoother for those adjusted for the human JND.

5.1.2 Labels

Labels were automatically generated nonsense words. The vowels a,e,i,o,u, plus some diphthongs and all consonants except x and q combined to form 500 consonant-vowel syllables. These were combined to produce a large number of two-syllable words. Any incidental real words were removed. The order of appearance of the words was the same for all experiments. The order was set to maximise the average edit distance between any adjacent labels. This meant that new labels presented to the participants would be relatively distinct, on average, from those already seen.

5.2 Procedure

The experiment proceeds in three parts: The Naming Game, a mutual exclusivity test and an elicitation of the participant's representation of the meaning space. Participants are seated at separate computer terminals. There may be two or more participants. Participants are paired randomly at each trial. Each participant has an avatar - an image of a character that represents them - so that participants can recognise who they are communicating with. Figure 5 shows a mock-up of the interface participants see.

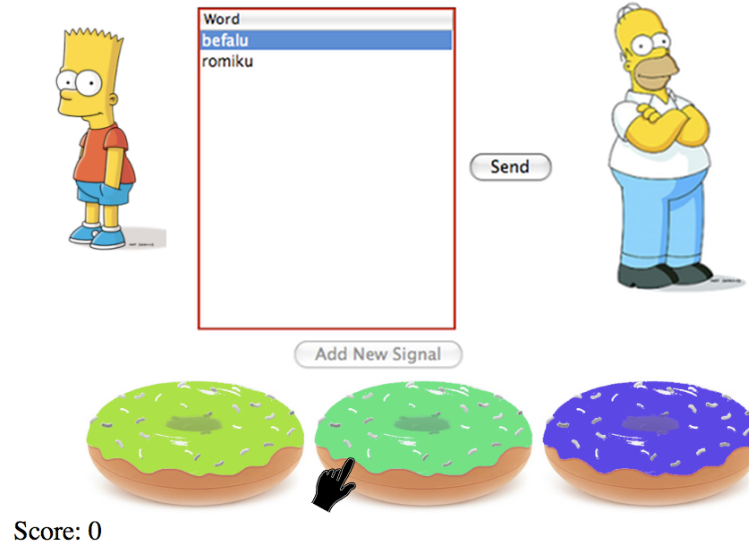


Figure 5: Screen-shot of the naming game interface. The participant's own avatar is on the right, their partner is shown on the left. Three coloured stimuli appear above. A hand symbol indicates the target to the speaker.

5.2.1 Naming Game

A single trial proceeds as follows:

1. Two individuals are chosen at random to be the speaker and the listener.
2. They both see M coloured squares. The perceptual distance between any pair of colours is at least greater than the human Just Noticeable Difference. One of the squares is selected as a topic (by the computer).
3. The speaker is presented with a list of labels with which they can label the colour. The speaker can also request that a new word be added to the list. The speaker selects a label and this is transmitted to the listener.
4. The listener receives the label and must guess which colour the speaker was referring to.
5. The speaker and listener receive feedback on whether they were correct or incorrect.
6. If the listener is correct, the word is added to the list of words which the listener can use in future rounds (if it is chosen to be a speaker).

The procedure above constitutes one trial. There are many trials in an experiment with the role of the speaker and listener being assigned randomly for each trial. In experiments with more than two people, who is paired with whom may also be randomised.

The experiment continues until a set number of trials have been completed.

5.2.2 Mutual exclusivity test

One final set of trials administers a mutual exclusivity test. Participants are told that they will meet an alien who they must try to understand. All participants are put in the role of a listener, communicating with a novel 'alien' avatar. This avatar is controlled by the computer and this test involves no real communication. They are shown a context with a familiar stimuli and a novel stimuli (either a colour in a part of the spectrum reserved for this test or a multi-coloured/patterned stimuli). The alien transmits a novel label. The participants must choose which stimuli the alien is referring to. Participants also indicate their confidence about their choice, rated on a 7 point Likert scale.

5.2.3 Meaning space representation

At the end of the experiment, the participants are given representations of the meaning space (a continuous coloured strip) and asked to label it with the labels they have learned, indicating which range each label refers to. In order to facilitate the largest number of possible approaches, this part will be done with pen and paper.

5.3 Quantification

The major measures of the model are the communicative success rate, the number of labels an agent has and various measures of how well co-ordinated agents' labels are. Communicative success is straightforward to measure. However, in this experiment, access to the internal representation of the meaning space and its labelling is impossible. The fixed list of possible responses is used as a way of limiting the participant's internal representation. Participants have to make their decision to create a new label explicit in the experiment. A measure of the increase in labels over time will be a key measurement.

In addition, the representations of the meaning space which the participants create at the end of the experiment will be used as if they were full and accurate representations of the meaning space. Measurements such as bin packing depth and efficiency, number of labels recalled, average width of category and entropy efficiency can be calculated. There is an issue with colour calibration, but there is no easy solution.

Results from the mutual exclusivity test will be compared across network types. These include preference for pairing the novel label with a novel stimuli or a familiar stimuli and confidence ratings. It is predicted that participants in the complex network will be more likely to choose the familiar object and less confident about their choice.

6 Extensions

This paradigm can be extended to larger networks. Also, the avatars of the participants can be manipulated. This can be done to change the identity of a speaker or to give the impression that there are greater or fewer participants in the network than there really are. This will affect the apparent distribution of variation over speakers.

7 Pilot

A short pilot was run with four human participants. The aim was to test the functionality of the experiment program rather than to run a full experiment. The context size was three and d_{min} was set at 0.1, meaning that there would be a maximum of roughly ten distinguishable colour categories. The brightness and saturation of colours were set at 90%. The network was balanced, so that everyone played everyone else an equal number of times. The intention was to give each player an equal number of turns as speaker and listener. Due to a programming error, however, two participants only ever played one role (speaker or listener). Players completed 25 rounds in about 13 minutes.

The results are shown below in figures 6 and 7. Communicative success improves, but drops off again towards the end. I interpret this as the words beginning to compete with each other and the system turning. Given more trials, the communicative success should improve. Figure 7 shows the use of signals during the experiment. One can see that, as the experiment progresses (moving up the chart), particular signals are re-used to refer to the same range of colours (horizontal range). For instance, 'romiku' takes up the middle, with 'belafu' and 'racoca' to each side. Even though the listeners are not always interpreting them correctly, the situation looks promising as far as communicative success is concerned.

At the end of the experiment, I asked participants to label a colour spectrum with the labels they could recall. Participants couldn't recall all the labels. For the one label that participants did recall, the average distance between the colour associated with the signal (romiku) was 11% of the spectrum. This is reasonably good, considering that a maximally efficient category width is 10% of the spectrum. In subsequent experiments, the list of labels could be given to the participants to ease this task.

Although this pilot was generally successful, there is not enough data here to make a judgement with regards to whether the participants are accommodating more than one lexicon.

References

Long, F., Yang, Z., and Purves, D. (2006). Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):6013–6018. [3]

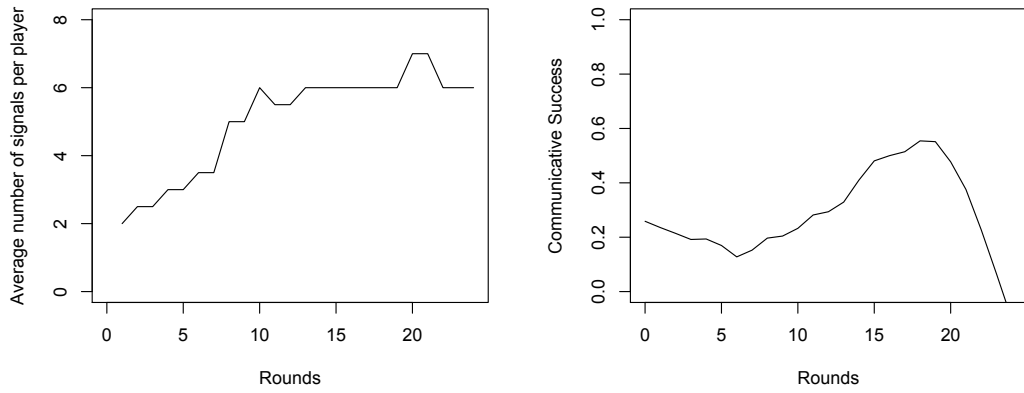


Figure 6: Average number of signals available to each participant (left). Average communicative success (right)

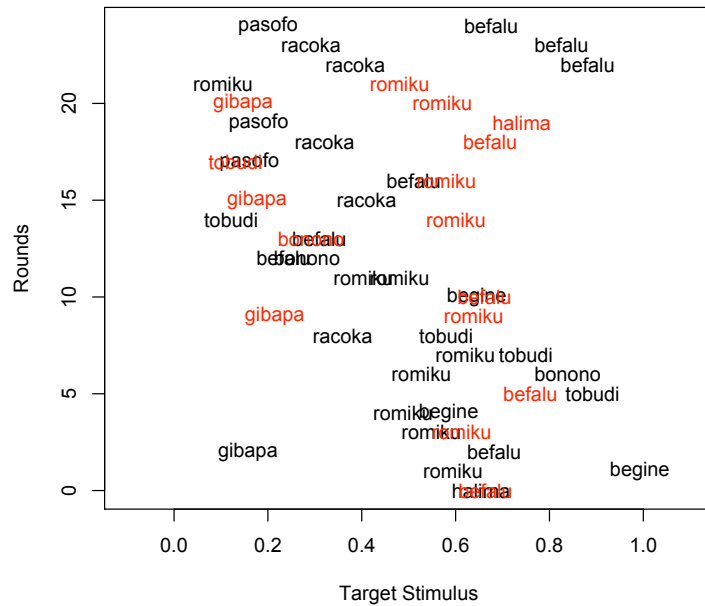


Figure 7: Signals used plotted by round and the value of the target colour. Signals in red indicate a successful communication.

BIBLIOGRAPHY

- Abrams, D. M. and Strogatz, S. H. (2003). Modelling the dynamics of language death. *Nature*, 424:900. [2, 10, 27, 58, 59, 60, 189, 190]
- Abutalebi, J., Annoni, J., Zimine, I., Pegna, A., Seghier, M., Lee-Jahnke, H., Lazeyras, F., Cappa, S., and Khateb, A. (2008). Language control and lexical competition in bilinguals: an event-related fMRI study. *Cerebral Cortex*, 18(7):1496–1505. [85]
- Altmann, E., Pierrehumbert, J., and Motter, A. (2011). Niche as a determinant of word fate in online groups. *PLOS ONE*, 6(5):e19009. [9]
- Andersen, R. (1982). Determining the linguistic attributes of language attrition. *The loss of language skills*, pages 83–118. [159, 161, 170]
- Andresen, J. (1990). *Linguistics in America 1769-1924*. Routledge. [4]
- Arikha, N. (2008). *Passions and tempers: A history of the humours*. Harper Perennial. [44]
- Arnold, D. (2012). Ken aboute language: Edinburgh students’ plans for school outreach to develop KAL and awareness of linguistics. In *Proceedings of the 2012 Edinburgh University Linguistics and English Language Postgraduate Conference*. [189]
- Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 21–30. Morgan Kaufmann Publishers Inc. [86]
- Au, T. K. and Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child Development*, 61(5):1474–1490. [165, 183]
- Auer, P. (1999). *Code-switching in conversation*. Routledge. [33]
- Ay, N., Flack, J., and Krakauer, D. (2007). Robustness and complexity co-constructed in multimodal signalling networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1479):441–447. [20]
- Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412. [164, 171]
- Baker, C. (2011). *Foundations of bilingual education and bilingualism*, volume 79. Channel View Books. [74]
- Baker, C. and Jones, S. P. (1998). *Encyclopedia of bilingualism and bilingual education*. Clevedon, UK: Multilingual Matters. [5, 74]
- Baker, M. (2003). Linguistic differences and language design. *Trends in cognitive sciences*, 7(8):349–353. [8, 14]
- Baldwin, J. (1896). A new factor in evolution. *The American Naturalist*, 30(354):441–451. [17]

- Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72(1):pp. 32–68. [36]
- Baron Cohen, S. (2004). Da ali g show, season 2 special features. James Bobin (Director). Bonus Material [Da Ali G show]. In Jeff Stilson (Producer): HBO. [49]
- Baronchelli, A., Gong, T., Puglisi, A., and Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107(6):2403–2407. [60]
- Batchelor, J. (1905). *An Ainu–English–Japanese Dictionary*. Tokyo:Methodist Publishing House, Ginza; London: Kegan Paul, Trench, Trubner Co. [34]
- Bates, E., Bretherton, I., and Snyder, L. (1991). *From first words to grammar: Individual differences and dissociable mechanisms*, volume 20. Cambridge University Press. [170]
- Bayes, T. and Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*, pages 370–418. [63]
- Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London. Online <http://www.cse.buffalo.edu/faculty/mbeal/thesis/index.html>. [86]
- Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, 59:1–26. [-]
- Bell, A. (1984). Language style as audience design. *Language in society*, 13(2):145–204. [9, 10, 129, 152]
- Berndt, R. M. (1959). The concept of ‘the tribe’ in the western desert of australia. *Oceania*, 30(2):81–107. [175]
- Bialystok, E. (1991). Metalinguistic dimensions of bilingual language proficiency. In Bialystok, E., editor, *Language processing in bilingual children*, pages 113–140. Cambridge University Press, Cambridge. [53]
- Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 65(4):229. [16, 189]
- Bialystok, E., Craik, F., Green, D., and Gollan, T. (2009). Bilingual minds. *Psychological Science in the Public Interest*, 10(3):89–129. [157]
- Bialystok, E. and Craik, F. I. M. (2010). Cognitive and Linguistic Processing in the Bilingual Mind. *Current Directions in Psychological Science*, 19(1):19–23. [2, 169, 178, 181, 189]
- Bialystok, E. and Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Developmental Science*, 7(3):325–339. [169]

- Birdsong, D. and Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of memory and language*, 44(2):235–249. [158, 161]
- Blanche, P. and Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39(3):313–338. [31]
- Blei, D., Griffiths, T., and Jordan, M. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7. [86]
- Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, volume 16, page 17. MIT Press. [86]
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press. [34]
- Bloomfield, L. (1927). Literate and illiterate speech. *American speech*, 2(10):432–439. [73]
- Bloomfield, L. (1933). *Language*. New York: Henry Holt. [i, 2, 6, 27]
- Blythe, R. A. (2009). Generic modes of consensus formation in stochastic language dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:P02059. [104]
- Blythe, R. A. (2012). Neutral evolution: A null model for language dynamics. *Advances in Complex Systems*, 15(1150015). [80]
- Boden, M. (2006). *Mind as machine: a history of cognitive science*, volume 1. Clarendon press. [156]
- Boeschoten, H. (1990). Asymmetrical code-switching in immigrant communities. In *Papers for the workshop constraints, conditions and models. Strasbourg: European Science Foundation Network on Code-Switching and Language Contact.*, pages 85–100. [9]
- Bolinger, D. (1968). Entailment and the meaning of structures. *Glossa*, 2(2):119–127. [15]
- Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. *Second language acquisition and the critical period hypothesis*, pages 133–159. [161]
- Boroditsky, L., Fuhrman, O., and McCormick, K. (2011). Do english and mandarin speakers think about time differently? *Cognition*, 118(1):123–129. [15]
- Bowern, C., Epps, P., Gray, R., Hill, J., Hunley, K., McConvell, P., and Zentz, J. (2011). Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS One*, 6(9):e25195. [38]
- Bowers, J. and Davis, C. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, 138(3):389. [122]
- Breal, M. (1893). On the canons of etymological investigation. *Transactions of the American Philological Association (1869-1896)*, 24:17–28. [4]

- Brendt, R. M. and Brendt, C. (1964). *World of the First Australians*. Chicago: University of Chicago Press. [34]
- Brigandt, I. (2003). Species pluralism does not imply species eliminativism. *Philosophy of Science*, 70(5):1305–1316. [46]
- Briscoe, T. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, pages 245–296. [73]
- Briscoe, T., editor (2002). *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge University Press. [60]
- Brojde, C., Ahmed, S., and Colunga, E. (2012). Bilingual and monolingual children attend to different cues when learning new words. *Frontiers in Psychology*, 3. [16, 24, 71, 83]
- Burkett, D. and Griffiths, T. (2010). Iterated learning of multiple languages from multiple teachers. In Smith, A., Schouwstra, M., de Boer, B., and Smith, K., editors, *The Evolution of Language: Proceedings of EvoLang 2010*, pages 58–65. [16, 55, 57, 65, 67, 69, 71, 72, 82, 88, 89, 104, 109, 202]
- Butler, Y. and Hakuta, K. (2004). Bilingualism and second language acquisition. *The handbook of bilingualism*, pages 114–144. [6, 7, 74, 157]
- Byers-Heinlein, K. and Werker, J. F. (2009a). Monolingual, bilingual, trilingual: infants’ language experience influences the development of a word-learning heuristic. *Developmental Science*, 12(5):815–823. [16, 24, 53, 71, 83, 183]
- Byers-Heinlein, K. and Werker, J. F. (2009b). Monolingual, bilingual, trilingual: infants’ language experience influences the development of a word-learning heuristic. *Developmental Science*, 12(5):815–823. [166]
- Calabria, M., Hernández, M., Branzi, F., and Costa, A. (2011). Qualitative differences between bilingual language control and executive control: evidence from task-switching. *Frontiers in Psychology*, 2. [84, 85, 168, 175]
- Call, J. and Tomasello, M. (2007). *The gestural communication of apes and monkeys*. Taylor & Francis Group/Lawrence Erlbaum Associates. [20, 22]
- Candolin, U. (2003). The use of multiple cues in mate choice. *Biological Reviews*, 78(4):575–595. [22]
- Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive neuropsychology*, 14(1):177–208. [84]
- Cartmill, E. and Byrne, R. (2010). Semantics of primate gestures: intentional meanings of orangutan gestures. *Animal cognition*, 13(6):793–804. [22]
- Castelló, X., Toivonen, R., Eguíluz, V. M., Loureiro-Porto, L., Saramäki, J., Kaski, K., and Miguel, M. S. (2008). Modelling language competition: bilingualism and complex social networks. In Smith, A., Smith, K., and Ferrer-Cancho, R., editors, *The evolution of language: Proceedings of the 7th International Conference (EVO LANG7), Barcelona 2008*. World Scientific Publishing Co. [10, 58, 59, 61]
- Chagnon, N. (1968). Yanomamo social organization and warfare. *War: The anthropology of armed conflict and aggression*, pages 109–159. [175]

- Chater, N. and Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*. [169]
- Cheney, D. and Seyfarth, R. (1988). Assessment of meaning and the detection of unreliable signals by vervet monkeys. *Animal Behaviour*, 36(2):477–486. [22]
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*, volume 119. MIT Press (MA). [1, 4, 27, 156, 157, 169]
- Chomsky, N. (1980). *Rules and representations*. Columbia University Press. [15, 46, 49, 52, 53, 156]
- Chomsky, N. (1982). *Noam Chomsky on the generative enterprise: a discussion with Riny Huybregts and Henk van Riemsdijk*. Foris Publications. [14]
- Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. Convergence. Praeger, New York. [156]
- Chomsky, N. (2000). *The architecture of language*. Oxford University Press. Online <http://www.chomsky.info/books/architecture01.htm>. [4, 27, 49, 52, 53, 154, 156]
- Chomsky, N. (2007). Symposium on margaret boden, mind as machine: A history of cognitive science. *Artificial Intelligence*, 171:1094–1103. [156]
- Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–508; discussion 509–58. [1, 50]
- Clahsen, H., Felser, C., Neubauer, K., Sato, M., and Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, 60(1):21–43. [44, 149]
- Clark, E. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum. [15]
- Clark, E. (1988). On the logic of contrast. *Journal of Child Language*, 15(02):317–335. [135]
- Colman, A. and Wilson, J. (2011). Antisocial personality disorder: An evolutionary game theory analysis. *Legal and Criminological Psychology*, 2(1):23–34. [187]
- Colomé, A. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of memory and language*, 45(4):721–736. [84]
- Cook, V. (1995). Multi-competence and the learning of many languages. *Language, Culture and curriculum*, 8(2):93–98. [157]
- Cook, V. (2010). multi-competence: knowledge of two or more languages in the same mind. Draft bits of Chapter for Chomsky's Universal Grammar 3rd edition, 2007. Online <http://homepage.ntlworld.com/vivian.c/Writings/Papers/multi-competence&UG.htm>. Retrieved 01/12/2010. [157]
- Cook, V. and Newson, M. (2007). *Chomsky's universal grammar: an introduction*. Wiley-Blackwell. [1, 4, 5, 156, 157]
- Cornish, H., Tamariz, M., and Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. *Language Learning*, 59:187–205. [20, 80, 83, 151, 181]

- Costa, A. (2005). Lexical access in bilingual production. *Handbook of bilingualism: Psycholinguistic approaches*, pages 308–325. [84]
- Costa, A. and Caramazza, A. (1999). Is lexical selection in bilingual speech production language-specific? further evidence from spanish–english and english–spanish bilinguals. *Bilingualism: Language and Cognition*, 2(03):231–244. [84]
- Costa, A., Caramazza, A., and Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1283. [84]
- Costa, A., Colomé, À., Gomez, O., and Sebastian-Galles, N. (2003). Another look at cross-language competition in bilingual speech production: Lexical and phonological factors. *Bilingualism: Language and Cognition*, 6(03):167–179. [84]
- Costa, A., Miozzo, M., and Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual’s two lexicons compete for selection? *Journal of Memory and Language*, 41(3):365–397. [84]
- Costa, A. and Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50(4):491–511. [71, 84, 85]
- Costa, A., Santesteban, M., and Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5):1057. [84]
- Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, 1984(46):49–70. [9]
- Cranshaw, A. (1998). *A study of Anglophone native and near-native linguistic and metalinguistic performance*. Universite de Montreal. [161]
- Craver, C. (2007). Explaining the brain. *Recherche*, 67:02. [44, 45]
- Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA. [15]
- Croft, W. (in press). *Language use and the evolution of languages*. The Language Phenomenon. Berlin: Springer. [26, 27, 36]
- Croft, W. A. (2003). Social evolution and language change. Unpublished manuscript. <http://www.unm.edu/~wcroft/Papers/SocLing.pdf>. [74, 161]
- Crystal, D. and Wang, F. (1997). *English as a global language*, volume 2. Cambridge University Press Cambridge. [5, 156]
- Cummins, J. (1976). *The Influence of Bilingualism on Cognitive Growth: A Synthesis of Research Findings and Explanatory Hypotheses*. Working Papers on Bilingualism, No. 9. Bilingual Education Project. [73]
- Curtis, A. (2007). *The Trap - What Happened to our Dream of Freedom: The Lonely Robot*. Adam Curtis (Director). In A. Curtis, L. Kelsall, S. Lambert (Producer): BBC. [121, 185, 187]

- Curtis, A. (2011). *The Trap - What Happened to our Dream of Freedom: The Lonely Robot*. Adam Curtis (Director). In D. Crossley-Holland & L. Kelsall (Producer): BBC. [187]
- Cutler, A. (2005). Lexical stress. *The handbook of speech perception*, pages 264–289. [41]
- Cutler, A., Mehler, J., Norris, D., and Segui, J. (1989). Limits on bilingualism. *Nature*, 340(6230):229–230. [41]
- Dabrowska, E. (2010). Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, 27(1):1–23. [6, 158, 159]
- Dahl, O. and Velupillai, V. (2011). The past tense. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. [125]
- Dale, P., Harlaar, N., Haworth, C., and Plomin, R. (2010). Two by two a twin study of second-language acquisition. *Psychological Science*, 21(5):635–640. [15]
- Darwin, C. and Ghiselin, M. (1874). *The descent of man*. Wiley Online Library. [50]
- Davies, A. (2003). *The native speaker: Myth and reality*, volume 38. Multilingual Matters Ltd. [6, 158]
- De Bot, K. (1992). A bilingual production model: Levelt’s speaking model adapted. *Applied linguistics*, 13:1–24. [84]
- De Groot, A. (2010). *Language and Cognition in Bilinguals and Multilinguals: An Introduction*. Psychology Press. [1, 4, 5, 6, 53, 83, 154, 156, 157, 158, 163, 166]
- De Houwer, A. (2007). Parental language input patterns and children’s bilingual use. *Applied Psycholinguistics*, 28(03):411–424. [60, 74, 175, 177, 201, 202, 203, 204]
- De Houwer, A. (2009). *Bilingual first language acquisition*. Multilingual Matters. [176]
- De Mejía, A. (2002). *Power, prestige, and bilingualism: International perspectives on elite bilingual education*, volume 35. Multilingual Matters Ltd. [75]
- de Saint-Exupéry, A. (1943). *The little prince*. Reynal & Hitchcock. [-]
- Deacon, T. (2010a). A role for relaxed selection in the evolution of the language capacity. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):9000–9006. [17, 23]
- Deacon, T. W. (2010b). A role for relaxed selection in the evolution of the language capacity. *Proceedings of the National Academy of Sciences*, 107(Supplement 2):9000–9006. [24]
- Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple bayesian models of language evolution generalise? *Journal of Theoretical Biology*, 259(3):552 – 561. [63, 64]

- Dell, G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283. [84]
- Dembeck, T. and Minnaard, L., editors (2012). *International Symposium on "Challenging the Myth of Monolingualism"*. [5]
- Dewaele, J. (1998). Lexical inventions: French interlanguage as L2 versus L3. *Applied Linguistics*, 19(4):471–490. [86]
- Dewaele, J. and Wei, L. (2012). Multilingualism, empathy and multicompetence. *International Journal of Multilingualism*, 1(15). [178]
- Diamond, J. (1991). *The Rise and Fall of the Third Chimpanzee*. Radius, London. [55, 62]
- Diaz, R. (1983). Thought and two languages: The impact of bilingualism on cognitive development. *Review of research in education*, 10:23–54. [174]
- Diebold, A. R. (1961). Incipient bilingualism. *Language*, 37(1):97–112. [6, 31, 32, 51]
- Doupe, A. J. and Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, 22(1):567–631. [19]
- Downs, A. (1957). An economic theory of political action in a democracy. *The Journal of Political Economy*, pages 135–150. [187]
- Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., and May, A. (2004). Neuroplasticity: changes in grey matter induced by training. *Nature*, 427(6972):311–312. [16]
- Drager, K. and Hay, J. (2006). Can you really believe your ears? the effect of stuffed toys on speech perception. In *New Zealand Language and Society Conference, Christchurch*. [32]
- Driesen, J. et al. (2011). Modelling vocabulary acquisition, adaptation and generalization in infants using adaptive bayesian pls. *Neurocomputing*, 74(11):1874–1882. [86]
- Dryer, M. (1992). The Greenbergian word order correlations. *Language*, pages 81–138. [90]
- Dunbar, R. (1993). Coevolution of neocortical size, group-size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–694. [24]
- Dunbar, R. (2003). The origin and subsequent evolution of language. In Christiansen, M. H. and Kirby, S., editors, *Language Evolution*, pages 219–234. Oxford University Press. [8, 14]
- Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82. [5, 8]
- Edwards, D. and Christophersen, H. (2011). Bilingualism, literacy and meta-linguistic awareness in preschool children. *British Journal of Developmental Psychology*, 6(3):235–244. [178]
- Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523. [115]

- Ellison, T. M. (2012). Topographic learning and the arbitrary, categorical lexicon. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 432–433. World Scientific. [137, 198]
- Elman, J., Diehl, R., and Buchwald, S. (1977). Perceptual switching in bilinguals. *The Journal of the acoustical Society of America*, 62:971. [32]
- Enfield, N. (2005). Areal linguistics and mainland southeast asia. *Annu. Rev. Anthropol.*, 34:181–206. [45]
- Eppler, W. (1993). Symbolic learning in connectionist production systems. In *Artificial Neural Nets and Genetic Algorithms*. [122]
- Erard, M. (2012). *Babel No More: The Search for the World’s Most Extraordinary Language Learners*. Free Press. [157]
- Ereshefsky, M. (1998). Species pluralism and anti-realism. *Philosophy of Science*, pages 103–120. [50]
- Ereshefsky, M. (2010). Darwins solution to the species problem. *Synthese*, 175(3):405–425. [46]
- Escudero, P. and Sharwood Smith, M. (2001). Reinventing the native speaker: or ‘what you never wanted to know about the native speaker so never dared to ask’. In Nizgorodcew, A. and Foster-Cohen, S., editors, *Euro SLA Yearbook 1*, volume 1, pages 275–286. John Benjamins Publishing Company. [6, 158, 159, 162, 163, 166]
- European Commission (2012). Special eurobarometer 386: Europeans and their languages. european comission. online http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_en.pdf. [5, 156]
- Evans, M. and Hatchwell, B. (1992a). An experimental study of male adornment in the scarlet-tufted malachite sunbird: I. the role of pectoral tufts in territorial defence. *Behavioral Ecology and Sociobiology*, 29(6):413–419. [22]
- Evans, M. and Hatchwell, B. (1992b). An experimental study of male adornment in the scarlet-tufted malachite sunbird: Ii. the role of the elongated tail in mate choice and experimental evidence for a handicap. *Behavioral Ecology and Sociobiology*, 29(6):421–427. [22]
- Evans, N. and Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492. [8, 15, 156]
- Fausey, C. and Boroditsky, L. (2008). English and spanish speakers remember causal agents differently. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, pages 64–70. [15, 83]
- Fausey, C. and Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic bulletin & review*, 18(1):150–157. [83]
- Fazly, A., Alishahi, A., and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*. [83]
- Ferdinand, V. and Zuidema, W. (2008a). *An experiment in iterated function learning*. ILLC Preprint Series (PP-2008-54), University of Amsterdam. [136]

- Ferdinand, V. and Zuidema, W. (2008b). *Language adapting to the brain: a study of a Bayesian iterated learning model*. ILLC Preprint Series (PP-2008-53), University of Amsterdam. [67]
- Ferdinand, V. and Zuidema, W. (2009). Thomas' Theorem meets Bayes' Rule: a model of the iterated learning of language. In Taatgen, N. and van Rijn, H., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society (Cogsci'09)*, pages 1786–1791. [1, 64, 65, 66, 67, 70, 90, 113, 115, 116, 117]
- Fernando, C., Valijarvi, R., and Goldstein, R. (2010). A model of the mechanisms of language extinction and revitalization strategies to save endangered languages. *Human biology*, 82(1):47–75. [58, 59, 60, 189, 190]
- Flack, J. and De Waal, F. (2007). Context modulates signal meaning in primate communication. *Proceedings of the National Academy of Sciences*, 104(5):1581. [24]
- Flower, T. (2011). Fork-tailed drongos use deceptive mimicked alarm calls to steal food. *Proceedings of the Royal Society B: Biological Sciences*, 278(1711):1548–1555. [21]
- Frank, M., Goodman, N., and Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–85. [16, 83, 86]
- Friston, K. and Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks*, 22(8):1093 – 1104. [123]
- Gafaranga, J. (2000). Medium repair versus other-language repair: Telling the medium of a bilingual conversation. *International Journal of Bilingualism*, 4:327–350. [7, 33]
- Gafaranga, J. (2007). 11. code-switching as a conversational strategy. *Handbook of multilingualism and multilingual communication*, 5:279. [17, 51]
- Gafaranga, J. (2008). *Talk in two languages*. Palgrave Macmillan. [7, 11, 17, 33, 45]
- Gal, S. (2009). Migration, minorities and multilingualism: Language ideologies in europe. In Mar-Molinero, C. and Stevenson, P., editors, *Language Ideologies, Policies and Practices: Language and the Future of Europe*, pages 13–28. Basingstoke: Palgrave. [156]
- Gao, J. and Johnson, M. (2008). A comparison of bayesian estimators for unsupervised hidden markov model POS taggers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 344–352. Association for Computational Linguistics. [86]
- Gardiner, C. (2009). *Stochastic methods. 4th Edition*. Springer. [174]
- Gathercole, S. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, 27(4):513. [173]
- Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press, USA. [27]
- Gelman, S. A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall, London. [194]

- Genesee, F., Hamers, J., Lambert, W., Mononen, L., Seitz, M., and Starck, R. (1978). Language processing in bilinguals. *Brain and language*, 5(1):1–12. [6]
- Godfrey-Smith, P. (2009). *Darwinian Populations and Natural Selection*. Oxford University Press. [82]
- Goedemans, R. and van der Hulst, H. (2011). Rhythm types. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. [41]
- Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism: Language and Cognition*, 6(01):1–15. [157]
- Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press. [15]
- Goldberg, A. (2009). Essentialism gives way to motivation. *Behavioral and Brain Sciences*, 32(05):455–456. [27]
- Goldwater, S. and Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics*, volume 45, page 744. [86]
- Gollan, T. and Kroll, J. (2001). Bilingual lexical access. In Rapp, B., editor, *The handbook of cognitive neuropsychology: What deficits reveal about the human mind*. Philadelphia, PA: Psychology Press. [84]
- Gomes, R., Welling, M., and Perona, P. (2008). Incremental learning of nonparametric bayesian mixture models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. [86]
- Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychol Sci*, 13(5):431–6. [179]
- Gong, T., Ke, J., Minett, J. W., and Wang, W. S.-Y. (2004). A computational framework to simulate the co-evolution of language and social structure. In *Artificial Life IX*, Boston, MA, U.S.A. [175]
- Gong, T., Puglisi, A., Loreto, V., and Wang, W.-Y. (2008). Conventionalization of linguistic categories under simple communicative constraints. In *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence)*. *IEEE Congress on*, pages 1686–1693. [60]
- Gong, T. and Wang, W.-Y. (2010). The roles of complex networks in linguistic categorization. *in press*. [60, 61]
- Gottlieb, G. (1971). *Development of species identification in birds: An inquiry into the prenatal determinants of perception*. U. Chicago Press. [22, 24]
- Gould, J. (1974). Honey bee communication. *Nature*, 252:300–301. [22]
- Gould, R. (1969). *Yiwara: foragers of the Australian desert*. Scribner. [-]
- Grassmann, S., Kaminski, J., and Tomasello, M. (in press). Differences in dogs’ and children’s understanding of acts of reference. *Animal Cognition*. [183]
- Green, C. and Bavelier, D. (2008). Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and Aging; Psychology and Aging*, 23(4):692. [16]

- Green, D. (1986). Control, activation, and resource: A framework and a model for the control of speech in bilinguals. *Brain and language*, 27(2):210–223. [84]
- Green, D. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and cognition*, 1(02):67–81. [84, 85]
- Greenberg, J. (1956). The measurement of linguistic diversity. *Language*, 32(1):109–115. [10, 29, 30, 31, 72]
- Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J., editor, *Universals of language*, pages 73–113. Cambridge, Mass. [8]
- Griffiths, P. (1997). *What emotions really are: The problem of psychological categories*. University of Chicago Press. [160]
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364. [122, 123]
- Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science: A Multidisciplinary Journal*, 31(3):441 – 480. [27, 41, 53, 54, 62, 63, 64, 65, 66, 68, 69, 71, 78, 79, 88, 89, 92, 96, 139, 169, 174, 175, 199, 202]
- Griffiths, T. L., Kalish, M. L., and Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363:3503–3514. [179]
- Griffiths, T. L., Lucas, C., Williams, J. J., and Kalish, M. L. (2009). Modeling human function learning with gaussian processes. *Advances in Neural Information Processing Systems*, 21. [194]
- Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4):334 – 384. [122]
- Grosjean, F. (1982). *Life with two languages: An introduction to bilingualism*. Harvard University Press. [74]
- Grosjean, F. (1985). The bilingual as a competent but specific speaker-hearer. *Journal of Multilingual & Multicultural Development*, 6(6):467–477. [7, 157]
- Grosjean, F. (1989). Neurolinguists, beware! the bilingual is not two monolinguals in one person. *Brain and language*, 36(1):3–15. [5, 17, 27, 83, 157]
- Grundy, P. (2007). Language evolution, pragmatic inference, and the use of english as a lingua franca. In Kecskes, I. and Horn, L. R., editors, *Explorations in pragmatics: linguistic, cognitive and intercultural aspects.*, pages 219–256. Berlin: Walter de Gruyter. [188]
- Gumperz, J. (1967). On the linguistic markers of bilingual communication. *Journal of social issues*, 23(2):48–57. [9]
- Gumperz, J. (1982). *Discourse strategies*, volume 1. Cambridge University Press. [2]

- Gumperz, J. and Hernandez-Chavez, E. (1972). Bilingualism, bidialectalism, and classroom interaction. In Cazden, C. B., Johns, V., and Hymes, D., editors, *Functions of language in the classroom*, pages 84–108. New York: Teachers College Press. [33]
- Gutiérrez-Clellen, V. and Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics*, 24(2):267–288. [176]
- Haffer, J. (1982). Systematik und taxonomie der larus argentatus-artengruppe. *Handbuch der Vögel Mitteleuropas*, 8:502–514. [50]
- Hagen, L. K. (2008). The bilingual brain: Human evolution and second language acquisition. *Evolutionary Psychology*, 6:43–63. [2, 14, 15]
- Haider, H. (1997). Typological implications of a directionality constraint on projections. In Alexiadou, A. and Hall, T. A., editors, *Studies on universal grammar and typological variation*, number 13, page 17. John Benjamins Pub Co. [90]
- Hakuta, K. (1989). Bilingualism and intelligence testing: An annotated bibliography. *Bilingual Research Group Working Papers*, Santa Cruz, California: University of California, Santa Cruz, Bilingual Research Group.(89-08). [157]
- Halmari, H. and Smith, W. (1994). Code-switching and register shift: Evidence from finnish-english child bilingual conversation. *Journal of Pragmatics*, 21(4):427–445. [10]
- Hamers, J. and Blanc, M. (2000). *Bilinguality and bilingualism*. Cambridge University Press. [6]
- Hamilton, W. and Zuk, M. (1982). Heritable true fitness and bright birds: a role for parasites? *Science*, 218(4570):384–387. [22]
- Han, W. and Huang, C. (2010). The forgotten treasure: Bilingualism and asian children’s emotional and behavioral health. *American journal of public health*, 100(5):831. [157]
- Hancin Bhatt, B. (2008). Second language phonology in optimality theory. *Phonology and Second Language Acquisition*. Philadelphia: John Benjamins. [4]
- Hansegård, N. (1968). *Tvåspråkighet eller halvspråkighet?* Stockholm: Aldus/Bonniers. [73]
- Hare, B., Plyusnina, I., Ignacio, N., Schepina, O., Stepika, A., Wrangham, R., and Trut, L. (2005). Social cognitive evolution in captive foxes is a correlated by-product of experimental domestication. *Current Biology*, 15(3):226–230. [24]
- Harper, D. (2012a). Online etymology dictionary: pace. <http://www.etymonline.com/index.php?term=pace> retrieved 28/06/2012. [37]
- Harper, D. (2012b). Online etymology dictionary: shriek. <http://www.etymonline.com/index.php?term=shriek> retrieved 28/06/2012. [38]
- Hatzidaki, A., Branigan, H., and Pickering, M. (2011). Co-activation of syntax in bilingual language production. *Cognitive psychology*, 62(2):123–150. [171]

- Haugen, E. (1953). *The Norwegian Language in America: A Study in Bilingual Behavior*, volume 1 of *The Bilingual Community*. Philadelphia: University of Pennsylvania Press. [6, 31]
- Haugen, E. (2009). Dialect, language, nation. *American anthropologist*, 68(4):922–935. [26, 31]
- Hauser, M. D. (1993). The evolution of nonhuman primate vocalizations: effects of phylogeny, body weight, and social context. *American Naturalist*, pages 528–542. [22]
- Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579. [5, 15, 19, 182]
- Hawkins, J. (1994). *A performance theory of order and constituency*, volume 73. Cambridge University Press. [90]
- Hay, J. and Drager, K. (2007). Sociophonetics. *Annu. Rev. Anthropol.*, 36:89–103. [33]
- Hay, J. and Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48(4):865–892. [32, 33]
- Hay, J. and Saffran, J. (2012). Rhythmic grouping biases constrain infant statistical learning. *Infancy*, 1(32). [41]
- Hayashi, M. and Matsuzawa, T. (2003). Cognitive development in object manipulation by infant chimpanzees. *Animal Cognition*, 6(4):225–233. [182]
- Healey, E. and Skarabela, B. (2009). Are children willing to accept two labels for one object?: A comparative study of mutual exclusivity in bilingual and monolingual children. In *Proceedings of the Child Language Seminar. University of Reading*. [16, 24, 53, 71, 83, 183]
- Helmbrecht, J. (2011). Politeness distinctions in pronouns. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. [125]
- Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L., and Feldman, M. W. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences*, 108:5154–5162. [15]
- Hermans, D., Bongaerts, T., de Bot, K., and Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, 1(3):213–229. [84]
- Hernández, M., Costa, A., Fuentes, L. J., Vivas, A. B., and Sebastián-Gallés, N., N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Bilingualism: Language and Cognition*, 13(3):315–325. [2, 53, 169, 181]
- Hill, J. (1978). Language contact systems and human adaptations. *Journal of Anthropological Research*, pages 1–26. [31, 34, 36, 48, 74]

- Hinton, G. and Nowlan, S. (1987). How learning can guide evolution. *Complex systems*, 1(1):495–502. [17]
- Hinton, L., Nichols, J., and Ohala, J. (2006). *Sound symbolism*. Cambridge University Press. [21, 137]
- Hirschfeld, L. (2008). The bilingual brain revisited: A comment on Hagen (2008). *Evolutionary Psychology*, 6(1):182–185. [15]
- Hirsh-Pasek, K., Golinkoff, R., and Hollich, G. (2000). An emergentist coalition model for word learning. In *Becoming a word learner: A debate on lexical acquisition*, pages 136–164. Oxford University Press. [71, 83]
- Hning, M., Vogl, U., and Moliner, O., editors (2012). *Standard Languages and Multilingualism in European History*. John Benjamins Publishing Company. [75]
- Hobaiter, C. and Byrne, R. (2011). The gestural repertoire of the wild chimpanzee. *Animal cognition*, 14(5):745–767. [20]
- Hoff, E. (2011). Studying children in bilingual environments. In *Research Methods in Child Language*, chapter 20, pages 300–316. Wiley-Blackwell. [167, 168]
- Hoffmann, C. (2001a). The status of trilingualism and bilingualism studies. In Cenoz, J., Jessner, U., and Hufeisen, editors, *Looking Beyond Second Language Acquisition: Studies in Tri- and Multilingualism*. Stauffenburg Verlag. [17]
- Hoffmann, C. (2001b). Towards a description of trilingual competence. *International journal of bilingualism*, 5(1):1–17. [83]
- Hoffmann, C. and Stavans, A. (2007). The evolution of trilingual codeswitching from infancy to school age: The shaping of trilingual competence through dynamic language dominance. *International Journal of Bilingualism*, 11(1):55–72. [17, 84]
- Holmes, W. and Mateo, J. (2007). Kin recognition in rodents: issues and evidence. *Rodent Societies: An Ecological & Evolutionary Perspective*, pages 216–228. [22]
- Holmes, W. and Sherman, P. (1983). Kin recognition in animals: The prevalence of nepotism among animals raises basic questions about how and why they distinguish relatives from unrelated individuals. *American Scientist*, 71(1):46–55. [22]
- Holt, B. and Benfer, Jr, R. A. (2000). Estimating missing data: an iterative regression approach. *Journal of Human Evolution*, 39(3):289–96. [198]
- Hopkins, W. and Leavens, D. (1998). Hand use and gestural communication in chimpanzees *Pan troglodytes*. *Journal of Comparative Psychology*, 112(1):95. [20]
- Hornberger, N. (2002). Multilingual language policies and the continua of bilit-eracy: An ecological approach. *Language Policy*, 1(1):27–51. [189]
- Hornberger, N. (2009). Multilingual education policy and practice: Ten certainties (grounded in indigenous experience). *Language Teaching*, 42(2):197–211. [189]

- Houston-Price, C., Caloghiris, Z., and Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, 15(2):125–150. [53, 71, 83, 183]
- Hultsch, H. and Todt, D. (1989). Context memorization in the song-learning of birds. *Naturwissenschaften*, 76(12):584–586. [23]
- Hultsch, H. and Todt, D. (1996). Discontinuous and incremental processes in the song learning of birds: evidence for a primer effect. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 179(3):291–299. [23]
- Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. *Logical issues in language acquisition*, pages 85–136. [4, 5]
- Hurford, J. R. (1991). The evolution of the critical period for language acquisition. *Cognition*, 40(3):159–201. [2, 14, 15, 16, 28, 32]
- Hurford, J. R. (1999). Desperanto. <http://www.lel.ed.ac.uk/~jim/desperanto.html> retrieved 12/10/2012. [181]
- Hurford, J. R. (2010). *The origins of meaning: Language in the Light of Evolution*. Studies in the Evolution of Language. Oxford University Press. [22]
- Hurford, J. R. and Dediu, D. (2009). Diversity in language, genes and the language faculty. In Botha, R. and Knight, C., editors, *The cradle of language*, Studies in the evolution of language, pages 163–184. Oxford University Press. [36, 38]
- Hyltenstam, K. and Abrahamsson, N. (2003). Maturation constraints in SLA. In Doughty, C. J. and Long, M. H., editors, *The handbook of second language acquisition*. Malden, MA: Blackwell. [6, 158, 161, 162]
- Ingram, D. (2003). English language policy in australia. In *Paper to the 2003 Summer International Conference of the Korea Association of Teachers of English*. [188]
- Inoue, S. and Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, 17(23):R1004–R1005. [183]
- Intrator, O. and Intrator, N. (2001). Interpreting neural-network results: a simulation study. *Computational statistics & data analysis*, 37(3):373–393. [122]
- Iriberry, N. and Uriarte, J. R. (2012). Minority language and the stability of bilingual equilibria. *Rationality and Society*. [189]
- Irvine, E. (2011). *Consciousness Science: A Science of What?* PhD thesis, The University of Edinburgh. [28, 44, 160]
- Isbell, L. A. and Young, T. (1996). The evolution of bipedalism in hominids and reduced group size in chimpanzees: alternative responses to decreasing resource availability. *Journal of Human Evolution*, 30:389–397(9). [24]
- Itkonen, E. (1996). Concerning the generative paradigm. *Journal of pragmatics*, 25(4):471–501. [156]
- Jackson, J. (1983). *The fish people: linguistic exogamy and Tukanoan identity in northwest Amazonia*. Number 39 in Cambridge studies in social anthropology. Cambridge University Press. [74, 175]

- Jacobs, M. (1937). Historic perspectives in indian languages of oregon and washington. *The Pacific Northwest Quarterly*, 28(1):55–74. [74]
- Jaeger, T. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, 59(4):434–446. [164, 171]
- Janson, T. and Schulman, R. (1983). Non-distinctive features and their use. *Journal of linguistics*, 19(02):321–336. [32]
- Jia, G. and Aaronson, D. (2003). A longitudinal study of chinese children and adolescents learning english in the united states. *Applied Psycholinguistics*, 24(01):131–161. [176]
- Johnson, J. and Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive psychology*, 21(1):60–99. [160]
- Jolly, A. (1966). Lemur social behavior and primate intelligence. *Science*, 153(3735):501–506. [20]
- Kagawa, H., Yamada, H., Lin, R., Mizuta, T., Hasegawa, T., and Okanoya, K. (2012). Ecological correlates of song complexity in white-rumped munias: The implication of relaxation of selection as a cause for signal variation in birdsong. *Interaction Studies*, 13(2):263–284. [23]
- Kahan, J., Setear, J., Bitzinger, M., Coleman, S., and Feinleib, J. (1992). *Developing games of local drug policy*. Rand Corporation. [187]
- Kalish, M. L., Griffiths, T. L., and Lewandowsky, S. (2007). Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychon Bull Rev*, 14(2):288–94. [115, 136, 138, 140, 194]
- Kaminski, J., Call, J., and Fischer, J. (2004). Word Learning in a Domestic Dog: Evidence for “Fast Mapping”. *Science*, 304(5677):1682–1683. [19, 183]
- Kanai, R. and Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, 12(4):231–242. [170]
- Kang, O. and Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, 28(4):441–456. [33, 160, 161]
- Kaplan, S. (2008). *Fixing fragile states: a new paradigm for development*. Greenwood Publishing Group. [188]
- Kaufman, T. and Thomason, S. (1988). Language contact, creolization and genetic linguistics. *Berkeley CA: University of California*. [47]
- Ke, J. (2004). *Self-organization and Language Evolution: System, Population and Individual*. PhD thesis, Department of Electronic Engineering, City University of Hong Kong. [54, 156]
- Kelley, L., Coe, R., Madden, J., and Healy, S. (2008). Vocal mimicry in songbirds. *Animal Behaviour*, 76(3):521–528. [21]
- Kemp, C. (2001). *Metalinguistic awareness in multilinguals: Implicit and explicit grammatical awareness and its relationship with language experience and language attainment*. PhD thesis, University of Edinburgh. [179]

- Kemp, C. (2007). Strategic processing in grammar learning: Do multilinguals use more strategies? *International Journal of Multilingualism*, 4(4):241–261. [179]
- Keysar, B., Hayakawa, S., and An, S. (2012). The foreign-language effect thinking in a foreign tongue reduces decision biases. *Psychological science*, 23(6):661–668. [83]
- Kindaichi, H. and Hirano, U. (1989). *The Japanese Language*. Tuttle Publishing. [35]
- Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In Knight, C., editor, *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pages 303–323. Cambridge University Press. [124]
- Kirby, S. (2001). Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity. *Evolutionary Computation, IEEE Transactions on*, 5(2):102–110. [169]
- Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10681–10686. [16, 20, 83, 137, 142, 169, 179]
- Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, 104(12):5241–5245. [1, 5, 53, 55, 63, 65, 66, 88, 89, 92, 96, 104, 124, 202]
- Kirby, S. and Hurford, J. R. (1997). Learning, culture and evolution in the origin of linguistic constraints. In *Fourth European Conference on Artificial Life*, pages 493–502. Citeseer. [80]
- Kirby, S. and Smith, K. (2010). Bayesian modelling code for simulating language course. [computer software]. [92]
- Kirby, S., Thompson, B., and Smith, K. (2012). Culture shapes the evolution of language. *Science*. [1, 5, 64, 65, 72, 104]
- Koda, H. (2012). Possible use of heterospecific food-associated calls of macaques by sika deer for foraging efficiency. *Behavioural Processes*. [21]
- Kostoulas-Makrakis, N. (2001). *Language and Society. Basic Concepts (in Greek)*. Athens: Metexmio. [5]
- Kostoulas-Makrakis, N., Karantzola, E., and Athanassiadis, E. (2006). Attitudes toward bilingualism: the case of two greek islands. *Mediterranean Journal of Educational Studies*, 11(2):17. [34]
- Kovács, A. M. and Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science*, 31(5940):611 – 612. [16, 24, 71, 83]
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon. [161]
- Kuhl, P. K. and Miller, J. D. (1978). Speech perception by the chinchilla: identification functions for synthetic vot stimuli. *Journal of the Acoustical Society of America*, 63:905–917. [19]

- Kurihara, K. and Sato, T. (2006). Variational bayesian grammar induction for natural language. *Grammatical Inference: Algorithms and Applications*, pages 84–96. [86]
- La Heij, W. (2005). Selection processes in monolingual and bilingual lexical access. *Handbook of bilingualism: Psycholinguistic approaches*, pages 289–307. [84]
- Labov, W. (1963). The social motivation of a sound change. *Word*, 19:273–309. [146]
- Lado, R. (1957). *Linguistics across cultures*. Ann Arbor: University of Michigan Press. [161]
- Lakoff, G. (1990). *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press. [15]
- Lamb, W. (2011). Is there a future for regional dialects in scottish gaelic? In Millar, R. M., editor, *Proceedings of the 'Linguistic attrition; Linguistic creation' Symposium, University of Aberdeen*. [27, 31]
- Lambert, R. and Freed, B. (1982). *The Loss of Language Skills*. ERIC. [159]
- Lambert, W. (1975). Culture and language as factors in learning and education. In Wolfgang, A., editor, *Education of immigrant students*. Toronto: Ontario Institute for Studies in Education. [6]
- Lauchlan, F., Parisi, M., and Fadda, R. (2012). Bilingualism in sardinia and scotland: Exploring the cognitive benefits of speaking a minority language. *International Journal of Bilingualism*. [157, 189]
- Laycock, D. (1985). Tok pisin and the census. In Wurm, S. A. and Muhlhausler, P., editors, *The Handbook of Tok Pisin*, pages 223–231. Department of Linguistics, Research School of Pacific Studies, Australian National University. [75]
- Leavens, D. and Bard, K. (2011). Environmental influences on joint attention in great apes: Implications for human cognition. *Journal of Cognitive Education and Psychology*, 10(1):9–31. [22]
- Lee, C. and Lee, D. (1985). A simple on-line bin-packing algorithm. *Journal of the Association for Computing Machinery*, 32(3):562–572. [97]
- Lee, I. and Ramsey, S. (2000). *The Korean language*. State University of New York Press. [35]
- Lee, R. (1972). !kung spatial organization: An ecological and historical perspective. *Human Ecology*, 1(2):125–147. [48]
- Lehmann, W. (1993). *Theoretical bases of Indo-European linguistics*. Routledge London and New York. [27]
- Levelt, W. (1993). *Speaking: From intention to articulation*. The MIT Press. [84]
- Levelt, W. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, 98(23):13464. [84]
- Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–75. [84]

- Levinson, S. (2003). *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge University Press. [15]
- Levinson, S. (2006). On the human interaction engine. *Roots of Human Society*. [34]
- Levinson, S. and Gray, R. (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*. [5]
- Levy, R., Reali, F., and Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in neural information processing systems*, 21:937–944. [86]
- Liebers, D., De Knijff, P., Helbig, A., et al. (2004). The herring gull complex is not a ring species. *Proceedings of the Royal Society of London-B*, 271(1542):893–902. [50]
- Lieberson, S. and Dil, A. (1981). *Language diversity and language contact: essays*, volume 16. Stanford University Press. [29]
- Lincoln, P. (1979). Dual-lingualism: Passive bilingualism in action. *Te Reo*, (22):65–72. [74, 161]
- Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the united states. *Language in Society*, 31(03):419–441. [33]
- Line, P. (2010). Horizontal cultural transmission of language: Investigating iterated learning with human dyads. Master’s thesis, University of Edinburgh. [179]
- Lipowska, D. and Lipowski, A. (2011). Naming game on adaptive weighted networks. *Arxiv preprint arXiv:1107.3263*. [61]
- Little, H. (2011). The role of foreigner directed speech in the cultural transmission of language & the resulting effects on language typology. Master’s thesis, University of Edinburgh. [33]
- Lo, A. (2002). Codeswitching, speech community membership, and the construction of ethnic identity. *Journal of Sociolinguistics*, 3(4):461–479. [180]
- Long, M. (1990). Maturation constraints on language development. *Studies in second language acquisition*, 12(3):251–285. [158]
- Loreto, V., Baronchelli, A., and Puglisi, A. (2010). Mathematical modeling of language games. In *Evolution of Communication and Language in Embodied Agents*, number 3, pages 263–281. Springer Berlin Heidelberg. [60]
- Lucas, C. G. and Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, 34(1):113–147. [83]
- Luchtenberg, S. (2002). Bilingualism and bilingual education and their relationship to citizenship from a comparative german-australian perspective. *Intercultural Education*, 13(1):49–61. [5]
- Lupyan, G. and Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, 5(1):e8559. [44]

- Lust, B., Flynn, S., Blume, M., and Westbrooks, E. (2010). Constructing adequate language documentation for multifaceted cross-linguistic data: A case study from the virtual center for study of language acquisition. In Grenoble, L. A. and Furbee, L., editors, *Language Documentation. Practices and values.*, John Benjamins Publishing Company. [179]
- MacEachern, S. (2001). Setting the boundaries: Linguistics, ethnicity, colonialism, and archaeology south of lake chad. *Archaeology, Language, and History: Essays on Culture and Ethnicity. John Edward Terrell, ed*, pages 79–101. [31]
- Machery, E. (2009). *Doing without concepts*. Oxford University Press, USA. [160]
- MacIntyre, P., Noels, K., and Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language learning*, 47(2):265–287. [31, 32]
- MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge university press. [86]
- Mackey, W. (1965). Bilingual interference: Its analysis and measurement. *Journal of Communication*, 15(4):239–249. [180]
- Mackey, W. (2000). The description of bilingualism. *The bilingualism reader*, pages 26–54. [6, 7, 8, 188]
- MacSwan, J. (1999). *A minimalist approach to intrasentential code switching*. Routledge. [4]
- MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk, volume 2: The database (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates. [178]
- MacWhinney, B., editor (2010). *Computational models of child language learning*, number 37; Special Issue 03. *Journal of Child Language*. [41]
- Magrath, R. D., Pitcher, B. J., and Gardner, J. L. (2009). An avian eavesdropping network: alarm signal reliability and heterospecific response. *Behav. Ecol.*, 20(4):745–752. [22]
- Maguire, E., Gadian, D., Johnsrude, I., Good, C., Ashburner, J., Frackowiak, R., and Frith, C. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398. [16]
- Mainzer, K. (2004). *Thinking in complexity: the computational dynamics of matter, mind, and mankind*. Springer, Berlin, 4th rev. and enl. ed edition. [217]
- Markman, E. (1992). Constraints on word learning: Speculations about their nature, origins, and domain specificity. In Gunnar, M. R. and Maratsos, M. P., editors, *Modularity and constraints in language and cognition; Minnesota Symposium on Child Psychology, 25*, pages 59–101. Hillsdale, NJ: Erlbaum. [183]
- Markman, E. M. and Abelev, M. (2004). Word learning in dogs? *Trends in Cognitive Sciences*, 8(11):479 – 481. [183]
- Markman, E. M. and Wachtel, G. A. (1988). Children’s use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20:121–57. [83, 183]
- Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co. [122]

- Martello, S. and Toth, P. (1990). *Knapsack problems: algorithms and computer implementations*. England: John Wiley & Sons Ltd. [97]
- Martin-Jones, M. and Romaine, S. (1986). Semilingualism: A half-baked theory of communicative competence. *Applied Linguistics*, 7(1):26–38. [15, 73]
- Mather, E., Schafer, G., and Houston-Price, C. (2010). The impact of novel labels on visual processing during infancy. *British Journal of Developmental Psychology*. [183]
- Matsuzawa, T. (2012). Outgroup: The study of chimpanzees to know the human mind. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 227–233. World Scientific. [183]
- Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press. [50]
- Mayr, E. (1963). *Animal species and evolution*. Harvard University Press; London: Oxford University Press. [52]
- McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., and Smith, L. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356. [79, 86, 122, 123]
- McCrohon, L. (2012). Simulating the effects of cultural coadaptation on rates of endogenous and exogenous language change. In McCrohon, L., Fujimura, T., Fujita, K., Martin, R., Okanoya, K., Suzuki, R., and Yusa, N., editors, *Five Approaches to Language Evolution: Proceedings of the Workshops of the 9th International Conference on the Evolution of Language*. World Scientific. [84]
- McCue, J., Lust, B., Corson-Rikert, J., Lowe, B., Paulson, J., Steinhart, G., Webb, F., and Westbrook, E. (2007). Sger: Planning information infrastructure through a new library-research partnership [final report]. *Online <http://hdl.handle.net/1813/5240>*. Retrieved 12/10/2012. [179]
- McLaughlin, B. and Nayak, N. (1989). Processing a new language: Does knowing other languages make a difference? *Interlingual processes*, 1:5. [179]
- Medgyes, P. (1992). Native or non-native: who’s worth more? *ELT journal*, 46(4):340–349. [6, 158]
- Merriman, W. E. and Bowman, L. L. (1989). The mutual exclusivity bias in children’s word learning. *Monographs of the Society for Research in Child Development*, 54(3-4). [83]
- Merriman, W. E. and Kutlesic, V. (1993). Bilingual and monolingual children’s use of two lexical acquisition heuristics. *Applied Psycholinguistics*, 14(2):229–49. [53, 71, 83, 165, 183]
- Mertens, S. (2006). The easiest hard problem: Number partitioning. In Percus, A., Istrate, G., and Moore, C., editors, *Computational complexity and statistical physics*. Oxford University Press, USA. [97]
- Meuter, R. and Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language*, 40(1):25–40. [84]

- Meyerhoff, M. (2008). Bequia sweet/bequia is sweet: syntactic variation in a lesser-known variety of caribbean english. *English Today*, 24(01):33–40. [146]
- Minnaard, L., editor (2012). *International Symposium on “Challenging the Myth of Monolingualism”*. Leiden University, Netherlands. June 23, 2012. [156]
- Mintz, A. and DeRouen Jr, K. (2010). *Understanding foreign policy decision making*. Cambridge University Press. [187]
- Monod, A. (1970). Multilinguisme des indiens trumai du haut-xingu (brésil central). *Langages*, 5(18):78–94. [74]
- Morton, E. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, pages 855–869. [21]
- Mufwene, S. (2001). *The Ecology of Language Evolution*. Number Cambridge University Press in Cambridge Approaches to Language Contact. Cambridge University Press. [47, 50]
- Mufwene, S. S. (2005). Language evolution: The population genetics way. In Hauska, G., editor, *Gene, Sprachen, und ihre Evolution*, number Gene, Sprachen, und ihre Evolution. Universitaetsverlag Regensburg. [31, 36, 37, 59]
- Mulder, K. and Hulstijn, J. (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied linguistics*, 32(5):475–494. [6, 158, 159]
- Muysken, P. (1997). Media lingua. *Contact languages: a wider perspective*, 17:365. [149]
- Myers Scotton, C. (1983). The negotiation of identities in conversation: A theory of markedness and code choice. *International journal of the sociology of language*, 1983(44):115–136. [2]
- Myers Scotton, C. (2006). *Multiple voices: An introduction to bilingualism*. Wiley-Blackwell. [74]
- Myers Scotton, C. and Ury, W. (1977). Bilingual strategies: The social functions of code-switching. *International Journal of the Sociology of Language*, 1977(13):5–20. [2]
- Naseem, T., Barzilay, R., and Globerson, A. (2012). Selective sharing for multilingual dependency parsing. In *Proceedings of ACL, 2012*. [86]
- Naseem, T., Snyder, B., Eisenstein, J., and Barzilay, R. (2009). Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36(1):341–385. [86]
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49. [187]
- Nation, R. and McLaughlin, B. (1986). Novices and experts: An information processing approach to the good language learner problem’. *Applied Psycholinguistics*, 7(1):41–55. [179]
- Nayak, N., Hansen, N., Krueger, N., and McLaughlin, B. (1990). Language-learning strategies in monolingual and multilingual adults. *Language Learning*, 40(2):221–244. [179]

- Nemeth, J. (2011). The prehistory of evolutionary linguistics. Master's thesis, Central European University. [4, 13]
- Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of anthropological archaeology*, 17(4):354–374. [8, 55]
- Nettle, D. (1999a). *Linguistic Diversity*. Oxford: Oxford University Press. [8, 29, 34, 47, 50]
- Nettle, D. (1999b). Using social impact theory to simulate language change. *Lingua*, 108(2-3):95–117. [3, 75]
- Nettle, D. and Romaine, S. (2000). *Vanishing Voices: The Extinction of the World's Languages*. New York: Oxford University Press. [8]
- Newman, M. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133. [38]
- Newman, M. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582. [38]
- Newman, P. (2000). *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press, New Haven. [9]
- Newman, S. (2002). Developmental mechanisms: putting genes in their place. *Journal of biosciences*, 27(2):97–104. [17]
- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, 18(1):62–85. [32]
- Niyogi, P. and Berwick, R. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences*, 106(25):10124–9. [65, 66]
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., Nelson, M., et al. (2008). Genes mirror geography within Europe. *Nature*, 456(7218):98–101. [36]
- Nowak, M., Komarova, N., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291(5501):114–118. [73]
- Nowak, M. A. and Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 96(14):8028–8033. [60]
- Núñez, R. (2010). Meaning coordination, semantic construals, and bodily action: implication for language evolution. In *JAIST International seminar on the Emergence and Evolution of Linguistic Communication*. [24]
- Oda, R. and Masataka, N. (1996). Interspecific responses of ring-tailed lemurs to playback of antipredator alarm calls given by Verreaux's sifakas. *Ethology*, 102:441–453. [21]
- Odds and evens (n.d.). *Wikipedia*. Retrieved September 11, 2012 from http://en.wikipedia.org/wiki/Odds_and_evens. [-]
- Okanoya, K. (2004). Song syntax in Bengalese finches: proximate and ultimate analyses. *Advance in the Study of Behaviour*, 34:297–346. [23, 24]

- Okanoya, K. (2010). Biological preadaptations for language. In *Proceedings of the JAIST International Seminar on the Emergence and Evolution of Linguistic Communication, Kyoto, Japan*. [23]
- Okanoya, K. and Merker, B. (2007). Neural substrates for string-context mutual segmentation: a path to human language. *Emergence of Communication and Language*, pages 421–434. [24]
- Onishi, T. (2010). Analyzing dialectological distributions of Japanese. *Dialectologia: revista electrònica*, pages 123–135. [35]
- Ota, M., Hartsuiker, R., and Haywood, S. (2010). Is a fan always fun? phonological and orthographic effects in bilingual visual word recognition. *Language and speech*, 53(3):383–403. [171]
- Otsuji, E. and Pennycook, A. (2010). Metrolingualism: Fixity, fluidity and language in flux. *International Journal of Multilingualism*, 7(3):240–254. [5, 34, 158]
- Paliwala, A. B. (2012). Language in Papua New Guinea: the value of census data. *Journal of the Linguistic Society of Papua New Guinea*, 30:1–31. [5]
- Paradis, J. (2011). Individual differences in child English second language acquisition. *Linguistic Approaches to Bilingualism*, 1(3):213–237. [154, 171, 172, 174, 176]
- Parra, M., Hoff, E., and Core, C. (2011). Relations among language exposure, phonological memory, and language development in Spanish-English bilingually developing 2-year-olds. *Journal of experimental child psychology*, 108(1):113–125. [154, 173]
- Peal, E. and Lambert, W. (1962). The relation of bilingualism to intelligence. *Psychological Monographs: general and applied*, 76(27):1–23. [6, 74, 157]
- Pearson, B., Fernández, S., and Oller, D. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language learning*, 43(1):93–120. [16, 83]
- Pelucchi, B., Hay, J., and Saffran, J. (2009). Statistical learning in a natural language by 8-month-old infants. *Child development*, 80(3):674–685. [41]
- Penman, A. (2010). A real education. *The Guardian*, 04/09/2010 <http://www.guardian.co.uk/lifeandstyle/2010/sep/04/andrew-penman-schools-education>. [189]
- Pennycook, A. (2010). The monolingual myth. *Language on the Move*, August 10th, 2010. Retrieved 18/07/2012. <http://www.languageonthemove.com/recent-posts/the-monolingual-myth>. [158]
- Pepperberg, I. M. and Wilcox, S. E. (2000). Evidence for a form of mutual exclusivity during label acquisition by grey parrots (*Psittacus erithacus*)? *Journal of Comparative Psychology*, 114(3):219–231. [183]
- Perfors, A. and Navarro, D. (2011). Language evolution is shaped by the structure of the world: An iterated learning analysis. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society. [137, 139, 179, 195]

- Petitto, L. A. and Kovelman, I. (2003). The bilingual paradox: How signing-speaking bilingual children help us to resolve bilingual issues and teach us about the brain's mechanisms underlying all language acquisition. *Learning Languages*, 8(3):5–18. [3, 5, 14, 15, 157, 177]
- Pifer, A. (1979). *Bilingual Education and the Hispanic Challenge*. New York: Carnegie Corporation. [187]
- Piller, I. (2000). Language choice in bilingual, cross-cultural interpersonal communication. *Linguistik online*, 5(1):1–20. [59]
- Pilley, J. and Reid, A. (2011). Border collie comprehends object names as verbal referents. *Behavioural processes*, 86(2):184–195. [183]
- Pinker, S. and Bloom, P. (1990). Natural selection and natural language. *Behavioral and Brain Sciences*, 13(4):707–784. [14, 17]
- Pinker, S. and Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95(2):201–236. [19]
- Poplack, S. (1988). Contrasting patterns of code-switching in two communities. In Wande, E., Anward, J., Nordberg, B., Steensland, L., and Thelander, M., editors, *Aspects of Multilingualism: Proceedings from the Fourth Nordic Symposium on Bilingualism, 1984*, pages 215–244. New York: Mouton de Gruyter. [33]
- Poullisse, N. and Bongaerts, T. (1994). First language use in second language production. *Applied linguistics*, 15(1):36–57. [84]
- Preston, D. and Long, D. (2002). *Handbook of perceptual dialectology*, volume 2. John Benjamins Publishing Company. [35]
- Pruden, S., Hirsh-Pasek, K., Golinkoff, R., and Hennon, E. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child development*, 77(2):266–280. [71, 83]
- Puglisi, A., Baronchelli, A., and Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Science*, 105:7936–7940. [60]
- Quay, S. (2008). Dinner conversations with a trilingual two-year-old: Language socialization in a multilingual context. *First language*, 28(1):5–33. [17]
- Quillinan, J., Kirby, S., and Smith., K. (2010). Co-evolution of language and social network structure through cultural transmission. In Smith, A. D., Shoustra, M., de Boer, B., and Smith, K., editors, *Proceedings of the 8th International Conference on the Evolution of Language (EVOLANG 8)*, pages 475–476. World Scientific. [175]
- Radcliffe-Brown, A. (1930). The social organization of australian tribes. *Oceania*, 1(1):34–63. [50]
- Rainey, H., Zuberbühler, K., and Slater, P. J. B. (2004). Hornbills can distinguish between primate alarm calls. *Proc R Soc Lond B.*, 271(755–759.t). [21]
- Ready-Morfitt, D. (1991). The role and limitations of self-assessment in testing and research. Unpublished manuscript, University of Ottawa, Ottawa, Canada. [31]

- Reali, F. and Griffiths, T. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):429–436. [58, 70, 79, 80, 189]
- Redican, W. and Rosenblum, L. (1975). *Facial expressions in nonhuman primates*. Stanford Research Institute. [20]
- Ritchie, G. and Kirby, S. (2005). Selection, domestication, and the emergence of learned communication systems. In *Second International Symposium on the Emergence and Evolution of Linguistic Communication*. [24]
- Ritchie, G. and Kirby, S. (2007). A possible role for selective masking in the evolution of complex, learned communication systems. *Emergence of communication and language*, pages 387–401. [17]
- Roberts, G. (2010a). *Cooperation, Social Selection, and Language Change: An Experimental Investigation of Language Divergence*. PhD thesis, The University of Edinburgh. [132, 133]
- Roberts, G. (2010b). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, 11(1):138–59. [8, 14, 132, 133, 146]
- Roberts, S. (2011). Mutual exclusivity in the emergence of a shared lexicon: A pilot laboratory experiment. Technical report, University of Edinburgh. [12, 263]
- Roberts, S. (in press). Monolingual biases in simulations of cultural transmission. In Dignum, V., Dignum, F., Ferber, J., and Stratulat, T., editors, *Integrating Cultures*. Springer. [12, 250]
- Roberts, S. and Quillinan, J. (in press). The qhimp qhallenge: Visual working memory in chimps and humans. In McCrohon, L., editor, *Student Responses to EvoLang9*. World Scientific. [183]
- Roberts, S. and Winters, J. (2012). Constructing knowledge: Nomothetic approaches to language evolution. In McCrohon, L., Fujimura, T., Fujita, K., Martin, R., Okanoya, K., Suzuki, R., and Yusa, N., editors, *Five Approaches to Language Evolution: Proceedings of the Workshops of the 9th International Conference on the Evolution of Language*, pages 148–157. World Scientific. [44]
- Rock-paper-scissors (n.d.). *Wikipedia*. Retrieved September 11, 2012 from <http://en.wikipedia.org/wiki/Rock-paper-scissors>. [-]
- Roelofs, A. (1998). Lemma selection without inhibition of languages in bilingual speakers. *Bilingualism: Language and Cognition*, 1(02):94–95. [84]
- Roeper, T. (1999). Universal bilingualism. *Bilingualism: Language and Cognition*, 2(03):169–186. [4, 5, 7, 35, 156, 158]
- Rogers, T. and McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press. [123]
- Rohwer, S. (1982). The evolution of reliable and unreliable badges of fighting ability. *American Zoologist*, 22(3):531–546. [22, 24]
- Roth, P. (1969). *Portnoy's Complaint*. paperback (1971), Corgi. [-]

- Roux, F. and Trémoulet, M. (2002). Organization of language areas in bilingual patients: a cortical stimulation study. *Journal of Neurosurgery*, 97(4):857–64. [2, 41, 42, 43]
- Rubin, D. (2011). The power of prejudice in accent perception: Reverse linguistic stereotyping and its impact on listener judgments and decisions. In Levis, J. and LeVelle, K., editors, *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference*, pages 11–17. Ames, IA: Iowa State University. [33, 160, 161]
- Rubin, D., Healy, P., Zath, R., Gardiner, T., and Moore, C. (1997). Non-native physicians as message sources: Effects of accent and ethnicity on patients' responses to aids prevention counseling. *Health Communication*, 9:351–368. [33, 160]
- Rumelhart, D. E. and McClelland, J. L. (1985). On learning the past tenses of english verbs. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel distributed processing, volume 1*. Cambridge, MA: MIT Press. [123]
- Saffran, J. (2012). Out of the brains of babes: Domain-general learning mechanisms & domain-specific systems. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 529–531. World Scientific. [82]
- Saffran, J., Hauser, M. D., Seibel, R., Kapfhamer, J., Tsao, F., and Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, 107(2):479–500. [41]
- Saffran, J., Johnson, E., Aslin, R., and Newport, E. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52. [41, 179]
- Saffran, J., Pollak, S., Seibel, R., and Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105(3):669–680. [41]
- Sapir, E. (1921). *Language*. (revised 1970) London: Harcourt, Brace and World. [-]
- Satterfield, T. (1999). *Bilingual selection of syntactic knowledge: extending the principles and parameters approach*. Springer. [4]
- Scheele, A. F., Leseman, P. P. M., and Mayo, A. Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics*, 31(01):117–140. [205]
- Schleicher, A. (1863). *Die Darwinsche Theorie und die Sprachwissenschaft: Offenes Sendschreiben an Herrn Dr. Ernst Hückel*. Weimar, H. Bohlau. [50]
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied linguistics*, 11(2):129–158. [128]
- Schwartz, B. (1998). The second language instinct. *Lingua*, 106(1-4):133–160. [-]
- Scott-Phillips, T. C., Hurford, J. R., Roberts, G., and Roberts, S. (2010). Pragmatics not semantics as the basis for clause structure. In Smith, A. D. M., Smith, K., de Boer, B., and Schouwstra, M., editors, *The Evolution of Language: Proceedings of the 8th International Conference on the Evolution of Language*, pages 491–492. Singapore: World Scientific. [181]

- Scott-Phillips, T. C. and Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9):411–417. [80, 151, 181]
- Seliger, H. and Vago, R. (1991). *First language attrition*. Cambridge University Press. [159, 161, 170]
- Serratrice, L., A., S., Filiaci, F., and Baldo, M. (2011). Pronominal objects in english–italian and spanish–italian bilingual children. *Applied Psycholinguistics*, 1(1):1–27. [166, 167]
- Serratrice, L., Sorace, A., Filiaci, F., and Baldo, M. (2009). Bilingual children’s sensitivity to specificity and generality: Evidence from metalinguistic awareness. *Bilingualism: Language and Cognition*, 12(2):19. [53]
- Seyfarth, R., Cheney, D., and Marler, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, 210(4471):801–803. [22, 24]
- Seyfarth, R. M. and Cheney, D. L. (1990). The assessment by vervet monkeys of their own and another species’ alarm calls. *Animal Behaviour*, 40:754–764. [21]
- Shapiro, C. (1989). The theory of business strategy. *The RAND journal of economics*, pages 125–137. [187]
- Shibatani, M. (1990). *The languages of Japan*. Cambridge University Press. [34]
- Shimmura, I. (1904). *The Boundary Lines between the East and West Japanese on Grammar*. reprinted in 1993, Kyoto: Shimmura Izuru Memorial Foundation. [35]
- Shipton, M. (2008). Storm over attack on welsh medium schools. *Wales Online*, Dec 29 <http://www.walesonline.co.uk/news/wales-news/2008/12/29/storm-over-attack-on-welsh-medium-schools-91466-22564471/>. Retrieved 17/07/2012. [157, 189]
- Shtatland, E. and Barton, M. (1998). Information theory makes logistic regression special. In *Annual Conference of NorthEast SAS Users’ Group (neseg)*, pages 190–93. [194, 195]
- Siewierska, A. (2011). Gender distinctions in independent personal pronouns. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. [9]
- Silver, M. and Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. *Trends in cognitive sciences*, 13(11):488–495. [198]
- Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold. [170]
- Skutnabb-Kangas, T. (1981). *Bilingualism or not: The education of minorities*. Clevedon, England: Multilingual matters. [73]
- Slocombe, K., Kaller, T., Turman, L., Townsend, S., Papworth, S., Squibbs, P., and Zuberbühler, K. (2010). Production of food-associated calls in wild male chimpanzees is dependent on the composition of the audience. *Behavioral Ecology and Sociobiology*, 64(12):1959–1966. [22, 24]

- Slocombe, K. and Zuberbühler, K. (2006). Food-associated calls in chimpanzees: responses to food types or food preferences? *Animal Behaviour*, 72(5):989–999. [22, 24]
- Slocombe, K. and Zuberbühler, K. (2007). Chimpanzees modify recruitment screams as a function of audience composition. *Proceedings of the National Academy of Sciences*, 104(43):17228–17233. [22]
- Smith, J. and Price, G. (1973). The logic of animal conflict. *Nature*, 246:15. [187]
- Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. [1, 5, 65, 66, 67, 68, 70, 71, 78, 104]
- Smith, K. and Kirby, S. (2008a). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1509):3591–3603. [64]
- Smith, K. and Kirby, S. (2008b). Natural selection for communication favors the cultural evolution of linguistic structure. In *Proceedings of the 7th International Conference (EVOLANG7)*, pages 283–290. [67]
- Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–386. [123, 124, 169, 170]
- Smith, K., Smith, A. D. M., and Blythe, R. A. (2011). Cross-situational learning: an experimental study of word-learning mechanisms. *Cognitive Science*. [83, 86, 179, 180]
- Smith, K. and Thompson, B. (2012). Iterated learning in populations: Learning and evolving expectations about linguistic homogeneity. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 227–233. World Scientific. [53, 65, 67, 70, 71, 72, 76, 77, 82, 83, 104, 169]
- Smith, K. and Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449. [150]
- Smith, M. S. (1993). Input enhancement in instructed SLA. *Studies in second language acquisition*, 15(02):165–179. [128]
- Snyder, B., Naseem, T., Eisenstein, J., and Barzilay, R. (2008). Unsupervised multilingual learning for POS tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050. Association for Computational Linguistics. [86]
- Sober, E. (1980). Language and psychological reality: some reflections on chomsky’s rules and representations. *Linguistics and Philosophy*, 3(3):395–405. [26, 45, 49, 50, 52, 53, 157]
- Sober, E. (1984). *The nature of selection: evolutionary theory in philosophical focus*. Cambridge, MA: MIT Press. [14]
- Sober, E. and Wilson, D. (1999). *Unto others: The evolution and psychology of unselfish behavior*. Harvard University Press. [187]
- Sollis, A. (2010). Multilingual schools can still achieve impressive results. *The Guardian*, 22/09/2010 <http://www.guardian.co.uk/commentisfree/2010/sep/22/multilingual-schools-impressive-results>. [189]

- Soma, M., Hasegawa, M., and Okanoya, K. (2009). Song-learning strategies in the bengalese finch: do chicks choose tutors based on song complexity? *Animal Behaviour*, 78:1107–1113. [23]
- Sorace, A. (1993). Incomplete vs. divergent representations of unaccusativity in non native grammars of italian. *Second Language Research*, 9(1):22–47. [161]
- Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. *Language*, pages 859–890. [5, 36]
- Sorace, A. (2004). Native language attrition and developmental instability at the syntax-discourse interface: Data, interpretations and methods. *Bilingualism Language and Cognition*, 7(2):143–145. [159, 161, 170]
- Sorace, A. (2011a). Cognitive advantages of bilingualism: is there a “bilingual paradox”? In Valore, P., editor, *Multilingualism. Language, Power, and Knowledge*, pages 335–358. Pisa: Edistudio. [3, 182]
- Sorace, A. (2011b). Pinning down the concept of “interface” in bilingualism. *Linguistic Approaches to Bilingualism*, 1(33). [1, 2, 156, 166, 170, 173]
- Sorace, A., Serratrice, L., Filiaci, F., and Baldo, M. (2009). Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua*, 119(3):460–477. [161]
- Sorensen, A. (1967). Multilingualism in the Northwest Amazon. *American Anthropologist*, 69(6):670–684. [45, 74]
- Spolsky, B. (2004). *Language policy*. Cambridge University Press. [188]
- Spruit, M. R. (2006). Measuring syntactic variation in dutch dialects. *Literary and Linguistic Computing*, 21(4):493–506. [36]
- Stadler, K. (2009). Cultural transmission and inductive biases in populations of bayesian learners. Master’s thesis, University of Edinburgh. Online <http://hdl.handle.net/1842/3617>. [66]
- Stauffer, D., Castelló, X., Eguíluz, V. M., and Miguel, M. S. (2007). Microscopic Abrams-Strogatz model of language competition. *Physica A: Statistical Mechanics and its Applications*, 374(2):835 – 842. [58]
- Steels, L. (1996). Perceptually grounded meaning creation. In Tokoro, M., editor, *Proceedings of the International Conference on Multi-agent Systems*. MIT Press, Cambridge, MA. [60]
- Steels, L. and Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489. [10, 60]
- Steels, L. and McIntyre, A. (1999). Spatially distributed naming games. *Advances in complex systems*, 1(4):301–323. [61]
- Steinhart, G., Saylor, J., Albert, P., Alpi, K., Baxter, P., Brown, E., Chiang, K., Corson-Rikert, J., Hirtle, P., and Jenkins, K. (2008). Digital research data curation: Overview of issues, current activities, and opportunities for the cornell university library. online <http://hdl.handle.net/1813/10903> retrieved 12/10/2012. [179]

- Sternberg, D. (2006). *Implicit and Unaltered: A Sequential Learning Approach to Bilingualism*. PhD thesis, Cornell University. [6, 18]
- Sternberg, D. and Christiansen, M. (2006). The implications of bilingualism and multilingualism for potential evolved language mechanisms. In Cangelosi, A., Smith, A. D. M., and Smith, K., editors, *Proceedings of the 6th International Conference on the Evolution of Language*, pages 333–340. [2, 6, 14, 17, 18, 157]
- Street, J. and Dabrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua*, 120(8):2080–2094. [6, 158, 159]
- Sulik, J. (2012). Abductive inference and insight in the evolution of symbolic communication. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 547–548. World Scientific. [86]
- Takahasi, M. and Okanoya, K. (2010). Song learning in wild and domesticated strains of white-rumped munia, *Lonchura striata*, compared by cross-fostering procedures: Domestication increases song variability by decreasing strain-specific bias. *Ethology*. [24]
- Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Shopen, T., editor, *Language typology and syntactic description, Vol. III: Grammatical categories and the lexicon*, pages 57–149. Cambridge, UK: Cambridge University Press. [15]
- Tamariz, M., Cornish, H., Roberts, S., and Kirby, S. (2012). The effect of generation turnover and interlocutor negotiation on linguistic structure. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 555–556. World Scientific. [1, 179]
- Tan, R. and Fay, N. (2011). Cultural transmission in the laboratory: agent interaction improves the intergenerational transfer of information. *Evolution and Human Behavior*, 32(6):399–406. [179]
- Thomas, L. and Wareing, S. (1999). *Language, Society and Power*. London: Routledge. [5]
- Thomas, M. S. and Allport, A. (2000). Language switching costs in bilingual visual word recognition. *Journal of Memory and Language*, 43(1):44 – 66. [32, 164]
- Thomason, S. G. (2000). Linguistic areas and language history. In Gilbers, D., Nerbonne, J., and Schaeken, J., editors, *Languages in Contact*, pages 311–327. Amsterdam: Rodopi. [45]
- Thompson, B. (2010). Communicative accuracy in ‘multilingual’ populations. Technical report, University of Edinburgh. [72, 73, 75]
- Thompson, B. (2012). Iterated learning from multiple teachers: Learning and evolving expectations about linguistic homogeneity. In *Proceedings of the 2012 Edinburgh University Linguistics and English Language Postgraduate Conference*. [53, 65, 72, 78]

- Thompson, B., Smith, K., and Kirby, S. (2012). Cultural evolution renders linguistic nativism implausible. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9)*, pages 227–233. World Scientific. [1, 5, 53, 55, 65, 67, 72]
- Tibbetts, E. and Dale, J. (2004). A socially enforced signal of quality in a paper wasp. *Nature*, 432(7014):218–222. [22]
- Tindale, N. and Jones, R. (1974). *Aboriginal tribes of Australia: their terrain, environmental controls, distribution, limits, and proper names*. University of California Press Berkeley. [-]
- Toivonen, R., Onnela, J., Saramaki, J., Hyvonen, J., and Kaski, K. (2006). A model for social networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):851–860. [58]
- Tôjô, M. (1938). *Hôgen to hôgengaku*. Shun'yôdô Shoten. [35]
- Tôjô, M. (1954). *Nihon Hogengaku*. Tokyo: Yoshikawa Kobunkan. [35]
- Tokugawa, M. and Miyazima, T. (1977). *Ruigigo jiten*. Tokyo: Tokydo Shuppan. [35]
- Tomasello, M. (2006). Why don't apes point? In Enfield, N. J. and Levinson, S. C., editors, *Roots of Human Sociality: Culture, cognition and interaction*, pages 506–524. Oxford & New York: Berg. Reprinted in R. Eckhardt (Ed.), (2008) *Variation, Selection, Development: The evolutionary model of language change*. Berlin: Mouton. [22]
- Touretzky, D. and Hinton, G. (1988). A distributed connectionist production system. *Cognitive Science*, 12(3):423–466. [122]
- Treccani, B., Argyri, E., Sorace, A., and Della Sala, S. (2009). Spatial negative priming in bilingualism. *Psychonomic bulletin & review*, 16(2):320–327. [53, 167, 168, 181]
- Unsworth, S. (2011). Utrecht bilingual language exposure calculator. Unpublished manuscript, Utrecht University. [176]
- Unsworth, S. (in press). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: The case of dutch gender. *Bilingualism: Language and Cognition*. [172, 173]
- Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A., and Tsimpli, I. (in press). On the role of age of onset and input in early child bilingualism in greek and dutch. *Applied Psycholinguistics*. [172, 173]
- Valdés, G. and Figueroa, R. (1994). *Bilingualism and testing*. Ablex Publishing. [6]
- Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., and Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33):13273–8. [195]
- van Wilgenburg, E., Sulc, R., Shea, K., and Tsutsui, N. (2010). Deciphering the chemical basis of nestmate recognition. *Journal of Chemical Ecology*, 36:751–758. 10.1007/s10886-010-9812-4. [19, 20]

- van Wuijtswinkel, K. (1994). Critical period effects on the acquisition of grammatical competence in a second language. *Language Learning*, 42(2):217–248. [161]
- Voegelin, C. and Harris, Z. (1951). Methods for determining intelligibility among dialects of natural languages. *Proceedings of the American Philosophical Society*, 95(3):322–329. [34]
- Vogel, E. and Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, 17(2):171–176. [170]
- Vogt, P. (2005). Meaning development versus predefined meanings in language evolution models. In *In Pack-Kaelbling*, pages 1154–1159. [60]
- von Frisch, K. (1947). *The Dances with the Honey Bee*. Institute for the Study of Animal Behaviour. [22]
- Warner, W. (1937). *A black civilization*. Harper & Row. [34]
- Wei, L. and Milroy, L. (1995). Conversational code-switching in a chinese community in britain: A sequential analysis. *Journal of Pragmatics*, 23(3):281–299. [9]
- Weinreich, U. (1953). *Languages in contact*. The Hague: Mouton. [6]
- Werker, J. and Byers-Heinlein, K. (2008). Bilingualism in infancy: first steps in perception and comprehension. *Trends in cognitive sciences*, 12(4):144–151. [16, 83]
- White, L. and Genesee, F. (1996). How native is near-native? the issue of ultimate attainment in adult second language acquisition. *Second language research*, 12(3):233–265. [161]
- Whittaker, E. (1910). *A History of the Theories of Aether and Electricity from the Age of Descartes to the Close of the Nineteenth Century*. Longmans, Green and co. [44]
- Wichmann, S. and Holman, E. W. (2009). Population size and rates of language change. *Human Biology*, 81(2-3):259–74. [47]
- Wierzbicka, A. (1988). *The semantics of grammar*, volume 18. John Benjamins Pub Co. [135]
- Winford, D. (2003). *An introduction to contact linguistics*, volume 33. Wiley-Blackwell. [47]
- Winter, B. and Christiansen, M. H. (2012). Robustness as a design feature of speech communication. In Phillips, T. C., Tamariz, M., Cartmill, E. A., and Hurford, J. R., editors, *The Evolution of Language: Proceedings of the 9th International Conference (EVO LANG9)*, pages 384–391. World Scientific. [20]
- Winters, J. (2011). Crossing those curious parallels: Investigating the relationships between segment inventory size, demography and subsystem tradeoffs. Master’s thesis, Cardiff University. [10, 106]
- Wolff, E. (2000). Language and society. In Heine, B. and Nurse, D., editors, *African Languages - An Introduction*, 317. Cambridge University Press. [5, 156]

- Yallop, C. (1969). The Aljawara and their territory. *Oceania*, 39(3):187–197. [175]
- Yamagiwa, J. (1967). On dialect intelligibility in japan. *Anthropological Linguistics*, 9(1):1–17. [34]
- Yildiz, Y. (2012). *Beyond the Mother Tongue: The Postmonolingual Condition*. Fordham University Press. [5, 156]
- Zahavi, A. (1975). Mate selectiona selection for a handicap. *Journal of theoretical Biology*, 53(1):205–214. [22]
- Zuberbühler, K. (2000). Interspecies semantic communication in two forest primates. *Proceedings of the Royal Society B: Biological Sciences*, 8(267-713). [21, 22]