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.

Sociocultural determination of linguistic complexity

Mark Atkinson

A thesis submitted for the degree of Doctor of Philosophy

School of Philosophy, Psychology, and Language Sciences

University of Edinburgh

2016

Abstract

Languages evolve, adapting to pressures arising from their learning and use. As these pressures may be different in different sociocultural environments, non-linguistic factors relating to the group structure of the people who speak a language may influence the features of the language itself. Identifying such factors, and the mechanisms by which they operate, would account for some of the diversity seen in the complexity of different languages. This thesis considers two key hypotheses which connect group structure to complex language features and evaluates them experimentally.

Firstly, languages spoken by greater numbers of people are thought to be less morphologically complex than those employed by smaller groups. I assess two mechanisms by which group size could have such an effect: different degrees of variability in the linguistic input learners receive, and the effects of adult learning. Four experiments conclude that there is no evidence for different degrees of speaker input variability having any effect on the cross-generational transmission of complex morphology, and so no evidence for it being an explanation for the effect of population size on linguistic complexity. Three more experiments conclude that adult learning is a more likely mechanism, but that linking morphological simplification at the level of the individual to group-level characteristics of a language cannot be simply explained. Idiosyncratic simplifications of adult learners, when mixed with input from native speakers, may result in the linguistic input for subsequent learners being itself complex and variable, preventing simplified features from becoming more widespread. Native speaker accommodation, however, may be a key linking mechanism. Speakers of a more complex variant of a language simplify their language to facilitate communication with speakers of a simpler language. In doing so, they may increase the frequency of particular simplifications in the input of following learners.

Secondly, esoteric communication — that carried out by smaller groups in which large amounts of information is shared and in which adult learning is absent — may provide the circumstances necessary for the generation and maintenance of more complex features. I assess this in four experiments. Without a learnability pressure, esoteric communication illustrates how complexity can be maintained, but there is generally no evidence of how smaller groups or those with greater amounts of shared information would develop comparatively more complex features. Any observable differences in the complexity of the languages of different types of groups is eliminated through repeated interaction between group members. There is, however, some indication that the languages used by larger groups may be more transparent, and so easier for adult learners to understand.

Lay summary

Languages are not all equally complex, and it is thought that there is a relationship between how complex a language is and the number of people who speak it: the more people who speak a language, the simpler the grammar of that language tends to be. It is not immediately clear why this would be the case, however, and I use a series of language learning experiments to test two possibilities. First, I test a theory that when a language has more speakers, the input the children get from which to learn the language is more varied. This increased variability stops them learning the more complex elements of the grammar, and so the language simplifies. I find no evidence for this claim.

Secondly, I test an alternative proposal that languages can simplify due to adult learning. Languages with more speakers are also those with a greater proportion of non-native speakers, those who learned the language as adults. Compared to children, adults are poor at learning complex grammar, and so their learning may cause the languages to become simpler. I find some support for this claim. Experiment participants do learn simpler grammar than that of the language they are trying to learn, but it is not easy to explain how an individual's simplification would then spread to affect the language as a whole. I suggest that when native speakers interact with adult learners, they simplify their language to successfully communicate. These simplifications will then be present in the data from which following learners acquire the language, and so those simplifications can spread.

Even if adult learning does explain how languages can simplify, however, this does not explain where the complexity in the languages came from in the first place, and why all languages are not simpler than they are. I therefore consider a claim that particular types of small communities of speakers might communicate in ways which would lead their languages to become more complex. I generally find little evidence to support this theory, but there is some indication that the languages spoken by larger groups of people may be more transparent: they are easier for outsiders to understand and learn.

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 has been executed by myself, except where due acknowledgement is made in the text.

Mark Atkinson

Acknowledgements

My PhD was funded by a much-appreciated studentship from the Arts and Humanities Research Council (AH/K503010/1).

Thanks most of all to my supervisors, Kenny Smith and Simon Kirby, for all your support, encouragement, ideas, and tolerance of my enthusiasm for running lots of participants in the search of rather large p-values. It's been an absolute pleasure doing a PhD with you, and (apart from the day of the dodgy randomisation function) it's been great fun. Thank you.

Thanks also to all in the Centre for Language Evolution past and present for all your help and support over the last few years. In particular, for the invaluable chats which have helped guide my research and for sharing your time, data, and methodologies, thanks to Alan Nielsen, Bill Thompson, Carmen Saldana, Catriona Silvey, Hannah Cornish, Hannah Little, James Winters, Jenny Culbertson, Jon Carr, Justin Sulik, Kevin Stadler, Marieke Schouwstra, Marieke Woensdregt, Matt Spike, Mónica Tamariz, Seán Roberts, and Yasamin Motamedi. And to all of you who shared an office with me, especially if you had to pilot one or more of the far less polished versions of my experiments, thanks to those above and also to Amanda Cardoso, George Starling, Steph DeMarco, and Thijs Lubbers.

To Greg Mills, thanks for collaborating with me on the tangram experiments. To Márton Sóskuthy and Olga Fehér, thanks for your recordings and everything else Hungarian. To Barbora Skarabela, Lauren Hall-Lew, and Mits Ota, thanks for your additional advice and encouragement. To Christine Caldwell, thank you for your support with completing the final experiment at Stirling.

Finally, thanks to Nicky for everything, but especially for finding a way to balance our lives enough for this PhD to happen. And to Leyla, for all your unbalancing.

Contents

Introduction

Explaining linguistic diversity is considered one of the primary goals of linguistics. It may be possible to explain some of this diversity by considering the non-linguistic factors which have been proposed to influence language features. More specifically, determining the effects of the sociocultural properties of the groups of people who speak a language may explain how and why different languages have evolved to display different degrees of complexity.

In this thesis, I consider two claims that relate group structure to language complexity: languages which have a greater number of speakers are simpler than those spoken by fewer people; and languages used primarily for "esoteric" communication in societies of intimates are more complex than those used for "exoteric" communication between strangers. My aim is to assess these related proposals experimentally, and in doing so evaluate the candidate mechanisms by which number of speakers and type of communication could have such effects.

This thesis is made up of 5 chapters. Chapter 1 provides a background to the experimental work which follows, including a historical overview of this area of research, the ways in which linguistic complexity can be defined and quantified, and a description of the key sociocultural factors which have been proposed as determinants of linguistic complexity. The following three chapters then describe my experimental investigations into these claims. Chapters 2 and 3 consider the proposal that languages with a greater number of speakers are simpler, by assessing two candidate mechanisms by which number of speakers could have such an effect: the degree of speaker variability in a learner's linguistic input, and the effect of adult learning. Chapter 2 describes four experiments which assess the mechanism of speaker input variability, and finds no evidence to support the proposal that linguistic input being provided by a greater number of speakers results in complex morphology being more difficult to acquire and so less likely to survive cross-generational transmission. Chapter 3 then describes three experiments which assess the mechanisms of adult learning, and the proposal that, in having greater proportions of non-native speakers, languages with greater numbers of speakers have adapted to the learning needs and preferences of adults to a greater extent. Adult learning is found to simplify morphological systems, but the introduction of these simplifications into the input for subsequent learners fails to provide a complete explanation for how such individual-level simplifications could affect group-level language characteristics. Native speaker accommodation to non-natives appears to be a key linking mechanism.

Chapter 4 shifts the focus from how sociocultural factors which largely focus on language learning can reduce linguistic complexity, to how they can maintain and generate complex features through language use. It describes four experiments which investigate the effects of esoteric communication on the emergence of linguistic conventions. In particular, it focuses on the roles of social network size and density, and the amount of shared knowledge that exists between speakers. I find very limited evidence that more esoteric communication results in more complex language, although there is some indication that the languages of smaller groups is less transparent from the perspective of out-group members.

I conclude with a general discussion of the implications of this research in Chapter 5, including alternative explanations to the proposed sociocultural determinants I have assessed, and offer directions for future research.

1 Sociocultural determination of linguistic complexity

1.1 Introduction

Languages are products of cultural evolutionary processes, transmitted across generations of users via social learning. They have therefore been repeatedly subjected to the selectional pressures of human learning and use; they have "evolved" in that they have been shaped by these pressures (Croft, 2000; Christiansen and Chater, 2008; Smith and Kirby, 2008; Beckner et al., 2009). As these pressures may well be different in different physical, demographic and sociocultural environments, non-linguistic factors may systematically determine features of languages (Croft, 1995; Nettle, 1999a; Wray and Grace, 2007; Trudgill, 2011; Dale and Lupyan, 2012). Identifying such conditioning factors, or (probabilistic) *determinants*, will account for some of the great linguistic diversity observable across the languages of the world (Evans and Levinson, 2009). Of particular interest to an increasing number of researchers is the identification of determinants which affect the structural properties of languages, the non-linguistic factors which may influence the degree of complexity in a language (Trudgill, 2011; Nettle, 2012).

In this chapter, I give an overview of the previous work which has considered how non-linguistic factors may explain cross-linguistic variation of language features and complexity, before describing the specific research questions for the experimental work of the following chapters. In Section 1.2, I give a historical overview of the field, and explain why it has been largely neglected for the best part of a century. In Section 1.3, I discuss what is typically meant by linguistic complexity, including different approaches to defining and quantifying it, before clarifying the position I take in this thesis. Section 1.4 provides a background in describing the previous work which has considered the non-grammatical features of language, and how the natural environment, human biology and sociocultural factors — those relating to a group's social structure, culture, and the interaction between the two — may account for crosslinguistic phonological and lexical diversity. I then focus on the determination of grammatical features: the core interest of this thesis's focus of the sociocultural determination of linguistic complexity. I then consider the determinants proposed to influence linguistic complexity in Section 1.5 and the evidence which supports them, before setting out the specific proposals I test in Chapters 2 to 4 in Section 1.6.

1.2 A historical perspective

> "[V]irtually all linguists today would agree that there is no hope of correlating a language's gross grammatical properties with sociocultural facts about its speakers."
>
> Newmeyer (2002, p. 361)

For most of the last century, it has been widely believed or argued that there was no relationship between the structure of a language (or any other features of language, for that matter, Nettle, 2012) and any characteristics of the natural environment, the genes of its speakers, or the sociocultural environment it was spoken in (Nettle, 2012). This was primarily due to two widespread assumptions: that language features (grammatical or otherwise) were arbitrary and so could not have been moulded by a type of society or culture, and by uniformitarianism: the prevailing belief that all languages were (necessarily) equally complex (Sapir, 1912; Kaye, 1989; Nettle, 1999b, 2012; Newmeyer, 2002; Deutscher, 2009; Sampson, 2009; Lupyan and Dale, 2010). Theories which offered sufficiently defined and interesting explanations for how linguistic features could be influenced by sociocultural factors were few and far between. Where they did exist, they were often overly specific to a particular language, supported by only anecdotal evidence, or unfalsifiable (Trudgill, 2011; Nettle, 2012). When more general, cross-linguistic claims were made, they often lacked credible explanatory mechanisms for how proposed determinants could have a described effect. There was also a lack of comparable datasets to support hypotheses, leading to an over-emphasis on counterexamples offered by the sceptical, who failed to consider probabilistic, rather than absolute, influences (Trudgill, 2011; Nettle, 2012).

Uniformitarianism stemmed from an argument that all languages are equally complex as they all use finite sets of elements to create and communicate (theoretically, at least) unbounded sets of meanings (Sampson, 2009; Trudgill, 2011; Nettle, 2012). Sampson (2009, p. 2) traces this view of functional equality across languages back to Hockett (1958, p. 180-1):

> "...impressionistically it would seem that the total grammatical complexity of any language, containing both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically. Fox, with a more complex morphology than English, thus ought to have a somewhat simpler syntax; this is the case."

Though Hockett's (1958) statement is somewhat qualified, his view is nevertheless a strong one (Sampson, 2009): a change in the complexity at one level of linguistic analysis will necessarily cause a decrease in complexity elsewhere. This view then largely became axiomatic, aided by ideological motives of descriptive linguistics and the rise of generativism. The idealogical motive was a response to a popular view that "primitive" societies would speak "primitive" languages. Many, at least as far back as Edward Sapir, were keen to point out that this was not the case (Sampson, 2009; Trudgill, 2011; Nettle, 2012). For generativist linguists, however, language became something unrelated to culture: the differences between languages lost their significance as phenomena of meaningful interest, and the focus on an innate cognitive system relevant to language acquisition was such that any meaningful variation in linguistic complexity was untenable (Sampson, 2009). Uniformitarianism, then, became an "urban legend" (Deutscher, 2009), and did not come under the level of scrutiny it should have done.

Crucially, the assumption of complexity invariance is flawed. Even accepting that all languages are capable of conveying the same, infinite set of semantic information, it does not follow that the encoding of this information is necessarily equally complex in every language (Nettle, 2012). There is also no reason why, as proposed, the sum of morphological and syntactic complexity would equal some constant. A verb in Archi, to take Sampson's (2009) example, can inflect into 1.5 million contrastive forms. Even if it could be quantified, it would be difficult to argue that a language with relatively simple morphology, such as English, has sufficient syntactic complexity to compensate. Tests of cross-linguistic invariant complexity also fail. Dahl (2009) compares the related languages of Elfdalian and Swedish, concluding that while Elfdalian is substantially more complex morphologically, there is very little evidence that Swedish is substantially more complex in some other domain. Nichols (2009) also attempts to compare the overall linguistic complexity of a set of languages. With a sample of 68 languages, and considering phonology, morphology, syntax, and lexis, she concludes that the evidence is more in favour of a probabilistic distribution of complexity, rather than there being some unifying constant. Further problems for uniformitarianism include different degrees of complexity within variants of the same language (Szmrecsany and Kortmann, 2009), among speakers of the same variant (Chipere, 2009), and over an individual speaker's lifetime (Sampson, 2009).[1] Finally, the very notion of a specific, quantifiable "overall complexity" measure is also flawed. Even if the various subdomains of a language could each be quantified, the complexity of each could not reasonably be summed to form some overall figure due to substantial differences in the phenomena they quantify. Therefore for two languages to be considered indisputably equal in complexity, they would have to be equally complex in each and every subdomain, which is not feasible (Deutscher, 2009).

Despite the opposition, there have been proposals of non-linguistic influences on language features for some time. Trudgill (e.g. 2011), for example, has notably argued for sociolinguistic effects on typology since the 1970s, Perkins (1992) has claimed an effect of (non-linguistic) cultural complexity on deixis, Thurston (e.g. 1994) has claimed that there are fundamental differences in the complexity of languages depending on whether they are primarily used for in-group communication or for interaction with outsiders, and McWhorter (e.g. 2005) has long argued that there is an effect of a language's age. With complexity invariance being less widely accepted and the existence of more comparable datasets, a number of studies are now more persuasively demonstrating that relationships between non-linguistic factors and language structure do exist (Nettle, 2012).

Before considering how non-linguistic factors could increase or decrease the complexity of a language, however, I focus on the different definitions of linguistic complexity and how it may be quantified.

1.3 Linguistic complexity

Defining linguistic complexity is far from trivial, and finding an all-purpose and uncontroversial definition is very possibly unachievable. Nevertheless, the notion of one language being more complex than another is clearly not a meaningless concept (McWhorter, 2005; Deutscher, 2009). In some cases, determining which is the more complex of equivalent subsets of two languages is even relatively straightforward. Consider past tense marking in Kikongo and Japanese, for example (adapted from McWhorter, 2005, p. 39):

English	Kikongo	Japanese
"I bought a goat (today)."	*Nsuumbidingí nkóombo.*	*Yagi o katta.*
"I bought a goat (yesterday)."	*Yásuumbidi nkóombo.*	*Yagi o katta.*
"I bought a goat (earlier)."	*Yasáumba nkóombo.*	*Yagi o katta.*
"I have bought a goat."	*Nsuumbidi nkóombo.*	*Yagi o katta.*
		(goat ACC bought)

Kikongo has four overt past tense distinctions compared to one past tense in Japanese. On this data, the grammatical past tense system of Kikongo would appear to be more complex than that of Japanese. Similarly, and with all other things being equal, a language with $n + 1$ grammatical distinctions or rules can be considered more complex than one with n (Dahl, 2004; McWhorter, 2005).

[1] Not only do children produce simpler sentences than adults, but there is some evidence that English-speaking 40 year-olds use more complex language than 30 year-olds, and that the over 60s use more complex language that those in their 40s and 50s (Sampson, 2009).

In many cases, however, ranking the complexity of a group of languages, even when only considering a single element, is far less trivial. McWhorter (2005, p. 43), illustrates the problem by considering plural marking in three different languages:

> "Complexity is certainly an ambiguous and malleable concept. For example, which plural marking strategy is more "complex": English, which marks plural only on the noun but with a marker that has three allomorphic variants; Swahili, which marks plurality redundantly on adjectives and nouns with a marker that varies according to several noun classes...; or French, which...marks plural via a plural allomorph of the determiner and redundantly on adjective and noun via a proclitic which is, however, expressed only when the root is vowel-initial...?"

The answer will largely depend on how you consider complexity. How, for example, do you relate the complexity of form to the complexity of the rules which determine those forms? Do you consider some objective measure of complexity, or equate it with how "difficult" the feature is for a language learner to acquire? If you focus on acquisition, are "complex" linguistic features the same for first language learners compared to second? (McWhorter, 2005). Do you consider what is complex for the speaker, or for the hearer? Discontinuous negation, for example, may be more complex for the speaker, but may ease the task of comprehension for the hearer (Mietsamo, 2009).

There are also ethnocentric and usage-assumption pitfalls to avoid. Language complexity has been overly assessed from a particularly Euro-centric viewpoint as opposed to a more objective one (Wray and Grace, 2007; McWhorter, 2009). There has also been a tendency to compare European languages with artificially-constructed, formal, and usually written, variants of non-European languages, rather than the more natural, and possibly very different, variants that native speakers would use spontaneously (Maas, 2009; Sampson, 2009). We also have to consider what we mean by a language. The linguistic complexity of an *individual's* language, for example, is variable over their lifespan (Sampson, 2009), and there is variability within the processing abilities of a group of native speakers of the same language: there is no "uniform grammatical competence" (Chipere, 2009).[2]

Despite these difficulties, there have been numerous attempts to define and quantify linguistic complexity. We can broadly group these into three (related) categories: L2 acquisition difficulty, feature-based, and information theoretic. We consider these in turn.

1.3.1 L2 acquisition difficulty

L2 acquisition difficulty, also referred to as "relative" or "outsider" complexity, is perhaps the most intuitive notion of language complexity for non-linguists, although it is also used within the field. Here, complexity is measured by the cost or difficulty of acquiring and processing a language for a non-native speaker (Dahl, 2009; Mietsamo, 2009; Szmrecsany and Kortmann, 2009).

The main advantages of this approach is that there is a direct relationship to the language user, with the potential to evaluate the complexity of different languages or language features through assessment of learner performance or psycholinguistic experiments. There are two main problems with this approach to complexity, however, particularly for the cross-linguistic notions

[2] This variation does not appear to be closely related to other factors, such as working memory, which may explain inter-speaking variability. There is instead some evidence for a more specific capability related to grammatical processing (Chipere, 2009).

of complexity we are concerned with in this thesis. What is costly or difficult for language users is not yet adequately understood (Mietsamo, 2009), and it is also too dependent on the individual user. With respect to acquisition, for example, complexity will be related to the language or languages already spoken by the learner, and the context of their acquisition. Native Swedish and English speakers may not both find the learning of Norwegian as "complex", nor might those who are exposed to the target language in a formal educational setting compared to those who are not. What is "difficult" depends very much on what an individual already knows (Dahl, 2004). Though there will strong correlations between some objective measure of complexity and the cost or efficiency from the point of view of the user (Hurford, 2012), it is best not to consider them equivalent at the outset (Sinnemäki, 2011).

Most (but not all; see, e.g., Szmrecsany and Kortmann, 2009) cross-linguistic analyses of complexity therefore avoid this relative approach to complexity, either due to the limitations mentioned above (McWhorter, 2005; Dahl, 2009; Mietsamo, 2009), or due to a more specific interest in what they may consider "strict linguistic structure" (Nichols, 2009, p. 111).

1.3.2 Feature-based structural complexity

A more objective, "absolute" (Dahl, 2009; Mietsamo, 2009), approach to quantifying the complexity of a language is to consider the number of categorical distinctions which are made within a given feature. In many ways, this is quite intuitive. Returning to our simple example on p. 4 and noting that Kikongo has 4 morphologically marked past tense distinctions compared to the single form in Japanese, we could crudely quantify the complexity of past tense marking as "4" for Kikongo and "1" for Japanese (reserving a score of "0" for languages with no past tense). Such an approach has the advantage of allowing direct cross-linguistic comparison using readily available datasets. Bentz and Winter (2013), for example, use data from the World Atlas of Language Structures (WALS) to investigate the distribution of case complexity. Feature 49A: Number of Cases (Iggesen, 2013) classifies 261 languages as either having no morphological case-marking, "borderline" case-marking, or 2, 3, 4, 5, 6-7, 8-9, or 10 or more case categories. Bentz and Winter (2013) compare this to the proportion of non-native speakers of a language to conclude that languages with greater proportions of adult learners have smaller case systems.

While this approach may be appropriate for analysis of a single feature, combining multiple features to achieve broader measures of complexity is problematic. Decisions have to be made about the relative contributions of each feature, which can be very difficult to justify (DeGraff, 2001; Mietsamo, 2009; Sampson, 2009). If the aim is to capture some *overall* measure of a language's complexity, there is also the substantial problem of capturing all aspects of its grammar (Deutscher, 2009; Mietsamo, 2009).

That said, complexity metrics can and have been devised to make useful typological observations and claims. In Nichols's (2009) assessment of uniformitarianism discussed above, she combined five individual complexity measures relating to phonology, morphological inflection, classification, syntax, and the lexicon, finding a negative correlation between community size and the complexity of a language. McWhorter (2005) uses a similar approach to illustrate his claim that creoles are less complex than other languages.

Considering a set of linguistic features together is also useful if the aim is to determine which features could be predicted by external variables. The effect of these variables can be assessed statistically, without having to derive an overall complexity measure for each language. The most comprehensive use of this approach which relates to language complexity was carried out by Lupyan and Dale (2010), which I return to below.

There are also some inherent disadvantages to such approaches, however. They make little allowance for the complexity of the rules which govern when a given category is applied for a given feature. When considering number of cases, for example, no distinction is made between a language which has 3 case markers applied equally frequently and one with 3 cases in which one is used the majority of the time. In the use of larger datasets, different methods of collecting the data for each language are largely ignored, and meaningful cross-linguistic comparison — e.g. assuming that the notion of a "word" or "case" in one language is directly comparable to a "word" or "case" in another — is taken too much for granted (Haspelmath, 2011).

1.3.3 Information-theoretic structural complexity

Information-theoretic approaches can more comprehensively quantify the complexity of language data, by considering complexity as being an inherent and objective property of that data or the system which produced it. Such approaches typically calculate the (minimum) amount of information necessary to encode that data, or the system which produces the data. In doing so, they can, for example, account for both the set of forms for a given linguistic feature, and the rules which govern the application of such forms (Clark, 2001; Brighton, 2003; Chater and Vitányi, 2003; Dahl, 2009).

Entropy (Shannon and Weaver, 1949), for example, allows analysis of reoccurring items of data with observed frequencies, such as phonemes and morphemes (Chater and Vitányi, 2003). For a set of observations S, entropy (in bits) is:

$$H(S) = -\sum_{s \in S} P(s) \log_2 P(s) \tag{1}$$

Taking the Kikongo data on p. 4 as a simple example, we can quantify how certain we are that a given past tense form will be used. If the four forms occur equally often, $H(S) = 2$. If one form occurs half the time and the others a sixth of the time each, $H(S) = 1.79$. On this basis, we have a measure of how certain we are that a given form will be used, and so have a quantifiable measure of how we can consider different sets of rules as having different degrees of complexity. This can then be extended to also consider the complexity of the forms, or the probabilities can be conditioned on some linguistic or extra-linguistic context.

Alternatively, Kolmogorov complexity can be used to quantify the complexity of any set of linguistic data. Unlike Shannon information theoretic measures, such as entropy, which associates probabilities with observed data items, Kolmogorov complexity can be applied to any object, and calculates the minimum number of bits from which that object can be reconstructed, i.e. the least redundant representation of the object, or the inverse of its compressibility (Clark, 2001; Brighton, 2003; Chater and Vitányi, 2003; Dahl, 2009; Mietsamo, 2009). To illustrate the approach and use Dahl's (2009) example, consider 3 strings, each of 6 characters: *hahaha*, *byebye*, and *pardon*. The first two strings can be compressed and represented as 3 x *ha* and 2 x *bye*, respectively, where as such an approach is not possible for *pardon*. *hahaha* can then be recoded as 3 characters and *byebye* as 4, while *pardon* can only be coded as 6. Assuming that each alphanumeric character can be minimally represented by the same number of bits, this suggests that *hahaha* has the lowest complexity, and *pardon* the greatest.

There are two issues with use of Kolmogorov for purposes such as the quantification of linguistic complexity. Firstly, it is dependent on the programming language used, though this is only up to some additive constant (Chater and Vitányi, 2003; Li and Vitányi, 1997).[3]. More problematically, it is not actually computable (Clark, 2001; Brighton, 2002). However, it can be approximated using the Minimum Description Length (MDL) principle (Rissanen, 1978, 1989; Brighton, 2002; Li and Vitányi, 1997). It states that the best description of some observed data is the one which minimises the lengths of both the description itself, and the length of the data under that description (Brighton, 2002; Li and Vitányi, 1997). This description will then be a representation of the data which is neither too simple, and so fails to capture the underlying characteristics of the data, nor one which overgeneralises and so capable of constructing erroneous data points (Brighton, 2002; Brighton et al., 2005).

Kolmogorov complexity and the MDL principle have been applied in research into the cultural evolution of language, by, for example, illustrating that a learner's bias for compression favours the emergence of compositional structure (Brighton, 2002, 2003; Brighton et al., 2005). It is not uncontroversially applied to topics such as linguistic complexity, however. Deutscher (2009) argues that as language cannot be fully described, then it can not be a suitable object for Kolmogorov complexity. Others argue that as it would find completely random data as maximally complex, it fails to capture a more intuitively language-relevant sense of complexity in which we may wish to only account for the structured pattens within an object, while excluding noise (Ay et al., 2008; Dahl, 2009; Mietsamo, 2009). A proposed solution is effective complexity (Gell-Mann, 1994). Here, a completely random object with no repeated segments has an effective complexity of 0, and so language data can be measured only in terms of its structure. The main criticism of such an approach is that deciding which parts of the object are those of interest and which are noise is too subjective, and so too dependent on the interests of whoever is applying it (McAllister, 2003).

1.3.4 Additional definitions of complexity

Information theoretic approaches could, theoretically at least, account for all linguistic behaviour (Brighton, 2003; Chater and Vitányi, 2003). They could therefore subsume the definitions below. I include these alternative definitions of complexity, however, as they are used in the literature relevant to sociocultural influences on cross-linguistic variation of language complexity, and I will revisit them later in this thesis.

Phonological complexity is alternatively considered the size of the phoneme inventory, cost or effort of phoneme production, phonotactic constraints or syllable structure (McWhorter, 2005; Nichols, 2009; Nettle, 2012). Semantic complexity refers to the tangibility (comparing concrete and abstract nouns for example), or the specificity and expressivity of a lexical item. "Labrador", for example, can be considered more semantically complex than "dog". *Abracadabra* would be particular complex, as "despite having no internal structural composition itself, [it] would require several words and some grammar" to define it in English (Wray and Grace, 2007, p. 569).

Bisang (e.g. 2009) also argues for a "hidden" notion of structural complexity, in addition

[3] For proof, see Li and Vitányi's (1997) chapter on the Invariance Theorem.

to the "overt" complexity discussed above. Overt complexity is explicit, observable language structure, the type of complexity most commonly referred to. By hidden complexity, he refers to the communicative complexity which is left to inference, rather than explicit encoding. Such complexity therefore "reflects economy: the structure of the language does not force a speaker to use a certain grammatical category if it can easily be inferred from context" (Bisang, 2009, p. 35). The possible consequence of ignoring hidden complexity is reaching a conclusion that a given language is more simple than it actually is: "simple" languages may actually not be so simple after all. As an example, Bisang (2009, p. 43) refers to the English sentence (from Lytinen, 1987, p. 305):

> *The stock cars raced by the spectators crowded into the stands at over 200 mph on the track at Indy.*

Though this is a relatively simple sentence in a language with relatively simple word order rules and few obligatory grammatical markers, it can be analysed in 156 different ways (apparently). Determining the intended analysis is left to the hearer and pragmatic inference (combined in this case with parsing complexity), reflecting a high degree of hidden complexity. Under this interpretation, the Japanese data may be more complex than the Kikongo in our example on p. 4.

Though Bisang (2009) accepts that quantifying hidden complexity is difficult, Ansaldo et al. (2015) demonstrate how it could be investigated with an fMRI study of syntactic processing in Mandarin speakers. They find that smaller clauses lead to greater amounts of semantic processing for the hearer. Bisang (2009) ultimately argues that only focusing on overt complexity means that only a limited range of language features are really being surveyed. Complexity as a whole is not being quantified.

1.3.5 Linguistic complexity for the purposes of this thesis

> "There are many questions about complexity which deserve linguists' full attention and best efforts: the evaluation of complexity in well-defined areas; the diachronic paths which lead to increases and decreases in complexity of particular domains; the investigation of possible links between complexity in particular domains and extra-linguistic factors, such as the size and structure of a society."
>
> Deutscher (2009, p. 251)

This thesis focuses on the third issue highlighted by Deutscher (2009) above: the effect of non-linguistic factors on linguistic complexity. In comparing different experimental conditions, I also touch on the second issue, in considering how language learning and use results in different levels of complexity in different social conditions: in Experiments 1 through 8, I investigate the causes of reductions in complexity; in Experiments 9 to 11, the emergence of linguistic conventions with potentially different levels of complexity. In evaluating the results of these studies, I do then quantify complexity in specifically well-defined areas.

As I am interested in the full range of hypotheses that previous investigations into sociocultural determination of linguistic complexity have generated, I deliberately avoid restricting myself to a particular metric or approach to defining complexity. Due to the methodologies of previous research, however, I do broadly consider overt, absolute complexity, as opposed to any notions of relative or hidden complexity.

In evaluating other researchers' hypotheses, I have designed experiments which assess changes to or the emergence of linguistic complexity in very specific areas, and with results which are comparable across experimental conditions. In doing so, I believe I largely avoid the problems of representativity and comparability (Mietsamo, 2009), in that I am not attempting to evaluate the complexity of all the elements of the languages encountered by the participants, nor am I comparing the complexity of one linguistic domain with another. The exact approach to defining complexity in the experiments primarily depends on what is most appropriate given the type of data. For the most part, information-theoretic measures such as entropy and mutual information prove the most appropriate and informative.

1.3.6 Function of language complexity

As I will discuss in Section 1.5, below, non-linguistic factors may determine the amount of variation in linguistic complexity observable across the languages of the world. Putting this variation aside for the moment, we can also consider why languages are so complex *in general.* Gil (2009), for example, argues that languages are generally much more structurally complex than strictly necessary. He argues that what he terms an "Isolating-Monocategorical-Associational Language", would completely suffice for the development and maintenance of human civilisation. Such languages have no word-internal structure, no distinct syntactic categories, and no semantic interpretations dependent on linguistic context. As all known languages are more complex than this (i.e. have additional structural properties) they are more complex than strictly necessary. Much of language structure is therefore functionless, and best considered as "the outcome of natural processes of self-organization whose motivation is largely or entirely system internal" (Gil, 2009, p. 33). Dahl (2004) also makes the point that, as many linguistic features are not present in all languages, they cannot strictly be necessary for communication.

This view has an overly idealised notion of natural communicative contexts, however. Trudgill (2010, p. 308), for example, relating redundancy specifically to inflectional morphology, notes that redundancy "seems to be necessary for successful communication, especially in less than perfect, i.e. normal, circumstances".

1.4 Determinants of non-grammaticial language features

Though I will focus on the determination of linguistic complexity, and am therefore primarily concerned with non-linguistic influences on cross-linguistic variation in grammatical structure, the determination of other language features is also relevant. We will later see how, for example, non-linguistic determination of the size of a phoneme set relates to morphological complexity (Nettle, 2012, and Chapter 2) and the effect of sound change on the formation of irregularities (Trudgill, 2011, and Section 4.1), while geographic features which isolate speech communities may influence the emergence of complex language features (Thurston, 1994, and Chapter 4). The determination of non-grammatical language features has also generally received a greater amount of attention, and so this research provides a background to the more specific claims assessed in this thesis.

The natural environment, human biology, and society and culture have each been proposed to affect phoneme inventories, phonotactic constraints, the presence of linguistic tone, and lexis. I consider these determinants and their effects in turn.

1.4.1 The natural environment

1.4.1.1 Phoneme inventory and phonotactic constraints

Cross-linguistic phonemic variation — including phoneme inventory size, presence or absence of particular phonemes, and phonotactic constraints — is largely presumed to be arbitrary (Ember and Ember, 2007a; Everett, 2013). A number of studies, however, have discussed potential links between the natural environment and variation in phonetic features. Much of this debate was originally driven by the research of Munroe and colleagues, and Ember and Ember, and is primarily based on noting correlations between temperature and phonemic variation, and then forming theories as to why such relationships would exist.

Munroe's group first analysed 53 languages and concluded that a warmer climate was linked to a language having a greater number of consonant-vowel syllables (Munroe and Silander, 1999). A later study, based on a sample of 60 languages, then linked warmer climates to a greater proportion of more sonorous phonemes (Fought et al., 2004). In both cases, they argued that the effect was due to the languages in warmer climates being adapted for distal communication.[4] Ember and Ember (2007a), in their reconsideration of Fought et al.'s (2004) data, are unconvinced by these conclusions. Rather than climate being the only possible determinant, they also put forward a number of other candidates, including the amount of vegetation in the area a language is spoken in, the amount of wind and rain, temperature range, terrain type and even the degree of sexual inhibition within the speech community. Munroe et al. (2009) later modify their claims regarding the quantity of consonant-vowel syllables by claiming that languages spoken in warmer climates have a greater number of sonorant consonants than those in cooler ones.

Atkinson (2011) also argues for an effect of geography on the phoneme inventory. From a sample of 504 languages, he finds a negative correlation between the size of its phoneme inventory, and the distance between where it is spoken and Africa. He argues that this is evidence for a common language root in Africa, with phoneme loss explained by successive founder effects: phonemic diversity, like genetic diversity, is reduced when a population is descended from a comparatively small group of individual settlers. This work has proved controversial, however, due to a lack of evidence that language divergence results in a reduction in phoneme inventories, a satisfying explanation for why this may happen, and concerns about autocorrelation (Nichols, 2009; Bybee, 2011; Dahl, 2011). Dahl (2011) also questions Atkinson's (2011) conclusions following a secondary analysis of his data.

Arguably, none of the studies above suggest a *direct* influence of the natural environment on phonetics (Everett, 2013). In Fought et al. (2004), for example, the natural environment may influence a society's culture, which in turn may influence their phoneme inventories. Everett (2013), however, finds a correlation between the elevation of the natural environment in which a language is spoken and the likelihood that that its phoneme set includes ejective phonemes. He proposes two explanations. The lower air pressure at higher altitudes may reduce the amount of physiological effort required for ejective production, and the use of ejectives may reduce the amount of water vapour lost to the speaker relative to other phonemes. Either (or both) of these explanations suggest an advantage for the speaker in having ejectives in their phoneme set. Everett et al. (2015) also claim that complex tonal languages are less likely to be spoken in environments with lower levels of humidity. They suggest that this is due to the inhalation

[4] As Fought et al. (2004, p. 29) note, this is dependent on assuming that "much of daily life" involves being outdoors to a greater extent in warmer climates compared to cooler.

of drier air affecting a speaker's ability to control vocal fold vibration and hence more carefully control pitch.

1.4.1.2 Lexis

For Trudgill (2011), there is an uncontroversial influence of the environment on lexical items. Norwegian has the word *ur* (masc.), meaning "rain clouds over the mountains", while other languages (e.g. English) lack a single-word equivalent. He argues that this is an effect of Norway's climate and topography, and the same reason that Paiute, spoken in south-west USA, has a wide range of words relating to desert topography (Trudgill, 2011). Similarly, other languages may have a greater quantity and more specific words for snow or sand, fish or reindeer, depending on the environments they are spoken in (Sapir, 1912; Nettle, 1999b; Trudgill, 2011). For Pullum (1991, p. 165), if this is true then it is "mundane and unremarkable", but argues that it may well not be:

> "[W]hen you come to think of it, Eskimos aren't really likely to be interested in snow. Snow in the traditional Eskimo hunter's life must be a kind of constantly assumed background, like sand on the beach. And even beach bums have only one word for sand."
>
> Pullum (1991, p. 166)

Regier et al. (2016) takes two samples of languages (one of 50 languages and one of 166) and find that those spoken in cooler climates are more likely to have separate words for "ice" and "snow", while those in warmer climates are more likely to have a single word for both. They conclude, therefore, that Pullum (1991) may be incorrect with respect to snow, and that variation in semantic categories may be dependent on whether or not such distinctions are useful to the language's speakers. It may be useful for Arctic communities to distinguish different types of snow, even if it is not for beach bums and different types of sand.

Changes in the environment can also cause observable lexical changes (Trudgill, 2011). As an example, there is the variant of English spoken on Tristan da Cunha, a remote and roughly circular island in the South Atlantic. The island was first inhabited in the early nineteenth century and the only settlement of note on the island, Edinburgh of the Seven Seas, is located on the north-west coast. By the 1920s, heading clockwise around the island from Edinburgh was trivially referred to as going *east*, but a walker would still be considered heading *east* for a whole circumnavigation of the island. Travelling what would have been referred to as "west" in the settlers' original variant of English from the south-east to the south-west of the island became referred to as *east* (Schreier, 2003, from Trudgill, 2011). Here, we see evidence of environment adaptation, with the language mirroring the locational system of other island-based languages. Manam, an Austronesian language spoken on the island of the same name off the north coast of New Guinea, for example, uses the terms *ata* and *awa*, equivalent to the *east* and *west* in Tristan da Cunha English, to effectively mean "clockwise" and "anticlockwise".

Munroe and Fought (2007), supported by Ember and Ember (2007b), make a claim relating the environment to word length. This may also be affected by the size of the phoneme inventory, with smaller inventories resulting in longer words (Trudgill, 2004a; Selten and Warglien, 2007; Sinnemäki, 2009). In reconsidering the data of Fought et al. (2004), they conclude that warmer climates result in shorter words. They go on to suggest that such an effect of climate on signal length may not be restricted to our species, drawing parallels with birdsong, which may be more characterised by short, loud bursts in more tropical regions (Marten and Marler, 1977).

1.4.2 Biological factors

1.4.2.1 Presence of linguistic tone

Genetic differences have also been claimed to have an impact on the phonological system of a language. The clearest example of this is where hereditary deafness leads to the use of a signed, rather than spoken, language (Meir et al., 2012). But Dediu and Ladd (2007) propose more subtle biological influences on language. They suggest that particular variants of ASPM (Abnormal Spindle-like Microcephally-associated) and MCPH1 (Microcephalin), two genes related to brain growth development, are determinants for a language possessing linguistic tone.[5] Through genetic and linguistic analysis of 49 populations, Dediu and Ladd (2007) note the correlation between an increased prevalence of individuals who have these genetic variants within a population, and a greater tendency for their language to have grammatical or lexical distinctions based on pitch. Their belief that this correlation is causal, due to these genetic variants strengthening an individual's weak bias to acquire or maintain linguistic tone, is controversial, however. Firstly, Bates et al. (2008) find no correlation between alleles of ASPM or MCPH1 and any traits relating to spoken or written language. Järvikivi et al. (2010) then argue that there is no clear-cut distinction between tonal and non-tonal languages, with both explainable by the same underlying cognitive mechanisms, and so the idea of genetic variation influencing the presence of linguistic tone is meaningless.

1.4.3 Sociocultural factors

1.4.3.1 Phoneme inventory

Following his research on Polynesian languages, Trudgill (2004a) suggests an influence of language contact on the size of the phoneme inventory. His proposals focus on the effects of post-critical period acquisition difficulties. He argues that language contact which leads to subsequent child language learning will lead to a larger phoneme set due to borrowing, as the learners will be able to acquire more complex sets. Contact which only involves adult learning, however, will result in a "medium-sized" phoneme inventory, which may better suit adult language acquisition as larger sets would be difficult for adults to acquire. Smaller sets would also increase word lengths, leading to words which adults would find difficult to acquire or find confusable. He also suggests that isolation, in involving only child acquisition, will not have a preference toward any phoneme inventory size as the learners will be able to acquire either longer words or a greater number of phonemes. Their languages may therefore have either very small or very large phoneme inventories.

He proposes similar effects of community size. Smaller communities, he argues, can tolerate phoneme sets of any size. Smaller sets could be the result of the speakers having "large amounts of shared information present" (Trudgill, 2004a, p. 317), reducing the need for grammatical redundancy (such as grammatical gender), and so reducing the need for longer words and more phonemes. Larger sets can also persist due to smaller societies being more linguistically conservative, resulting in more faithful cross-generation transmission of more complex (e.g.

[5] A variant of ASPM, more frequently found in Europe and the Middle East compared to Asia, is also claimed to have been involved in the invention of alphabetic writing. Frost (2008) argues that it may explain why ideographs are prevalent in the Far East and alphabets are favoured in writing systems elsewhere.

larger) phoneme inventories. Conversely, larger communities will disfavour smaller phoneme inventories, as a greater number of more distinctive phonemes may help avoid communicative difficulties. Though not commenting on phoneme set size specifically, Wray and Grace (2007) make a related claim that in smaller, more unified groups of people, there are more likely to be what they term “unusual” sounds or “difficult” sound combinations. They also put this down to the effects of faithful transmission in smaller groups: as the learning will be primarily, or exclusively, be done by children, sound combinations which would be difficult for non-native speakers to acquire will be able to persist.

In a sample of 428 languages, Pericliev (2004) finds no evidence to support Trudgill’s (2004a) hypothesis that larger populations will have medium-sized inventories. Trudgill’s (2004b) response is that population size should not be considered as a predictive variable on its own, but needs to be considered in relation to the community’s social network structure and contact with speakers of other languages. Hay and Bauer (2007), however, do find evidence of a correlation between population size and number of phonemes. Both in reanalysis of Pericliev’s (2004) data (originally used only to test for medium-sized phoneme sets in languages spoken by more people) and in their own sample of 216 languages, they find that languages spoken by a greater number of people have larger phoneme inventories.

Ember and Ember (2007a) also propose causal links between sociocultural factors and the proportion of more sonorant phonemes in a language. The argue that societies with more limited infant-holding traditions will have more sonorant phonemes, to facilitate more distal communication between child and primary caregiver (cf. the claims of Munroe and Silander, 1999, and Fought et al., 2004, in Section 1.4.1.1). They also relate more sonorant phonemes to higher levels of expressiveness or lower sexual inhibition in a society, by appealing to the openness of the mouth and amount of nasality when singing as an indication of an individual’s sexual permissiveness, for example. More evidence is clearly needed to support such claims.

1.4.3.2 Lexis

Similar to effects of the environment (see Section 1.4.1.2, above), sociocultural factors can also affect lexis, with the number and specificity of ways of referring to a semantic grouping being highly dependent on an individual’s and society’s needs and experiences (Sapir, 1912; Trudgill, 2011). Almost trivially, we could note the differences in the amount of vocabulary relating to sheep between sheep-farming and non-sheep-farming communities (Trudgill, 2011). While such observations are uncontroversial (Trudgill, 2011), there are also more interesting proposals. Wray and Grace (2007) argue that in smaller populations, individuals are more likely to share interests, occupations and experiences. This will lead to a greater number of “highly-specific” lexical items. Wray and Grace (2007) do not expand on this specific point, but, as with the size of sheep vocabulary, this could be due to the usefulness of making more semantic distinctions for a larger proportion of the group (I return to this claim when considering semantic complexity in Experiments 9 and 10 in Chapter 4).

1.5 Sociocultural determinants of grammatical complexity

More interesting and challenging to explain than many of the relationships discussed so far is the influence of non-linguistic determinants on grammatical structure (Trudgill, 2011). To my knowledge, there are no claims which specifically and directly relate any aspect of the natural environment or human biology to systematic variation in grammatical complexity. Nevertheless, a number of theories have been proposed for how sociocultural factors can affect the structure of a language, and so determine the degree of a language's complexity. These are primarily related to the social structure of its speakers.

Thurston (e.g. 1994), Wray and Grace (2007) and Trudgill (e.g. 2011) consider how linguistic complexity is influenced by social factors as a result of different communicative contexts and pressures. Trudgill originally made such proposals in 1977 (Trudgill, 2011), and he identifies five interdependent features of an esoteric group:[6] community size, social network structure, social stability, contact, and the extent to which information within a group is "shared". He predicts more complex languages will be spoken "in communities with the following constellation of societal features:

- low amounts of adult language contact
- high social stability
- small size
- dense social networks
- large amount of communally-shared information."

(Trudgill, 2011, p. 146)

Contact involving adults, he argues, results in simplifications, such as the regularisation of irregularities, and increases in lexical and morphological transparency, where semantic categories more clearly map to linguistic expressions. "Two times", for example, is more transparent than "twice", "eye-doctor" more than "optician", and "did go" more than "went" (Trudgill, 2011). There will also be loss of redundancy, both syntagmatic (repetition of information) and paradigmatic ("the morphological expression of grammatical categories", Trudgill, 2011, p. 22). Such simplifications, Trudgill (2011) hypothesises, are a result of adults finding irregularities, opaque forms, syntagmatic redundancy and greater numbers of morphological categories difficult to acquire. Conversely, contact involving child learning may lead to language complexification in the form of additive (rather than replacive) borrowing from one language to another as a result of bilingualism.[7] This increase in complexity will need a high degree of social stability to support it.

These predictions describe how languages can lose or maintain existing complex features, but do not explain the origin of such features (Thurston, 1994; Trudgill, 2011). Low-contact societies, however, are proposed to not only maintain linguistic complexity, but also increase it. This may be particularly the case in societies with fewer members, denser social networks, and large amounts of shared knowledge (Trudgill, 2011). Such conditions naturally nurture the development *mature phenomena* (Dahl, 2004): more complex linguistic features.

Thurston (1994) and Wray and Grace (2007) broadly support these proposals, seeing esoteric

[6] Trudgill (2011) does not generally use the terms "esoteric" and "exoteric" himself to describe different types of social group or communication, but I have adopted them when describing his hypotheses to highlight the overlap between his predictions and those of Thurston (1994) and Wray and Grace (2007).

[7] This separation of contact effects into those relating to adult and child contact is sometimes referred to as the "Trudgill insight" (e.g. Wright, 2012, 2013). Prior to this, there was more of a disagreement about the effects of contact on the presence of complex features.

communicative contexts[8] as breeding grounds for morphological complexity and irregularities,[9] opaque forms and idioms, and derivational constraints leading to a more suppletion. They also highlight esoteric communication leading to unusual sounds and more difficult sound combinations, and in the formation and maintenance of a greater number of semantically complex, highly-specific lexical items (as discussed in Sections 1.4.3.1 and 1.4.3.2). At the other extreme, there are exoteric contexts, with communication being used by larger groups, with a greater amount of interaction with unknown individuals and therefore more limited communally-shared information present, either due to the different interests and experiences of the different individuals, or due to through the interlocutors having less common ground due to limited past interactions. This will be characterised by more one-to-one relations between form and meaning, regularity, transparency, flexibility of expression and compositionality of signals.

Thurston (e.g 1994) also highlights the differences in speed of simplification and complexification. Exoteric pressures can stimulate rapid changes, and these will be simplifying ones. In the short term, this can be be an effect of adult learning difficulties resulting from contact. Even more immediately, this can be a result of in-group members adopting a simplified communicative strategy when interacting with strangers. Esoteric complexification effects, however, will be gradual. The languages of north-western New Britain, for example, while having similar morphosyntactic and semantic structures, vary vastly in terms of distinctions within a grammatical category and number of distinctions, which Thurston (1994) attributes to differences in their duration of isolation.

While Trudgill (2011), Thurston (1994), and Wray and Grace (2007) give specific and detailed examples from individual languages to support their hypotheses, analyses of larger language datasets also support some of their hypotheses. The most extensive of these studies was carried out by Lupyan and Dale (2010). They consider 28 structural features taken from the World Atlas of Language Structures (WALS) data, relating to morphological type, case system, verb morphology, agreement, possibility and evidentials, negation, plurality, interrogatives, tense, possession, aspect, mood, articles, demonstratives and pronouns. Through analysis of 2,236 languages, they conclude that a language's morphological complexity can be related to its number of speakers, the area it is spoken over, and the number and type of neighbouring languages. Larger languages, or those more widely spoken, tend to have less morphological complexity, while less widely spoken languages are marked by higher levels of grammatical complexity. Interdependent effects of language family and geography were controlled for. While the general correlation was observed whether population size, area, or linguistic contact were considered, population size was found to have the greatest predictive power.

These results replicate those of previous, smaller-scale studies. Nichols (2009), in analysis of 68 to 215 languages (depending on the particular language feature under investigation; for the measure of total complexity, her sample size was 130), concluded that smaller populations have more complex languages, with more extensive inflectional synthesis, classification systems

[8] A relationship between "esoteric" groups, "esoteric" communication and "esoteric" languages is not absolute, but they are closely connected. A group with a more esoteric social structure is more likely to engage in a greater proportion of esoteric (compared to exoteric) communication. Therefore the language (inasmuch as a language can be considered a stable unit for analysis) is then more likely to display features which have arisen from this greater use of esoteric communication (Thurston, 1994; Wray and Grace, 2007; Trudgill, 2011). As Wray and Grace (2007) in particular stress, though, all languages are likely to be used for both esoteric and exoteric communication to some extent, and the same language may be employed differently by different speech communities within the total population of speakers.

[9] The use of morphological strategies rather than lexical may is in itself likely to cause an increase in the number of irregular forms (Jackendoff, 1999).

(such as a greater number of noun classes), syntax (such as a greater number of basic word orders), phonology (such as more vowel quality distinctions), and lexis. She notes, however, that this result may simply be a chance geographical effect, as the correlation failed to hold within continents, though Lupyan and Dale's (2010) sample was both larger and controlled for such effects. Sinnemäki (2009) also concluded from a sample of 50 languages that there is a relationship between community size and linguistic complexity when considering core argument marking, such as the use of morphological strategies which mark subject and object.

For Lupyan and Dale (2010) (also Dale and Lupyan, 2012), this correlation between morphological complexity and population size is due to the effect of adult learning. Languages with greater numbers of speakers are also those with greater proportions of adult learners. Proposing that adults have difficulties with the acquisition of complex morphology, they argue that languages with simpler morphology have adapted to the learning needs and preferences of their non-native speakers. In having less morphological specifications, such languages will also be less redundant. Redundancy, for Lupyan and Dale (2010, Text S9), "refers to the degree that grammatically encoded information specifies something that can be readily extracted from context or pragmatics". They also illustrate how this redundancy can be specifically compared across languages. Using written translations of the Universal Declaration of Human Rights from 130 languages (all of which use the Roman alphabet, for comparison purposes),[10] they use a compression algorithm[11] to find a negative correlation between the compressibility ratio of the texts and the population of speakers of each language (Lupyan and Dale, 2010, Text S10; Table S2).

While suggesting that increased levels of redundancy hampers adult learning, Lupyan and Dale (2010) speculate that it may actually aid child learning. As increases in morphological complexity will increase redundancy, infants are provided with additional linguistic cues to supplement their relatively undeveloped abilities to use contextual information (Trueswell et al., 1999; Snedeker and Trueswell, 2004; Weighall, 2008). Simpler and more complex languages have therefore adapted to the needs of adult and child language acquisition, respectively. I discuss these proposals in more details in Chapter 3.

There is a lot of overlap in the hypotheses of Trudgill (2011) and Wray and Grace (2007) on the one hand (with support for the simplifying effects of language contact involving adult learning from, e.g., Dahl, 2004), and Lupyan and Dale (2010) (along with Nichols, 2009, and Sinnemäki, 2009) on the other. Both sets of accounts highlight group size as being an influence on language complexity, and both discuss the role of adult learning in simplification. A difference lies in the first camp viewing group size as only one of a set of interdependent factors which would likely determine linguistic complexity,[12] while the second demonstrates that population size alone is a predictor of (at least morphological) complexity, with adult learning being the explanatory mechanism. They could be reconciled by noting the likely interdependence of group size and the other factors, but they do appear to also place different emphases on the importance of different features of linguistic complexity, however. While Wray and Grace (2007), for example, stress opaque forms, irregularities and cases of suppletion as being representative of esoteric communication, as well as features which would increase syntagmatic and

[10] As mentioned in Section 1.3, note that translations of formal written texts may poorly reflect more natural, spontaneous speech, particularly in non-European languages (Maas, 2009).

[11] WinRAR, available at http://www.rarlab.com/.

[12] Trudgill (2004b, 2011) has been particularly keen to point out that the factors need to be considered together, often in response to empirical work which only considers a single factor and fails to find a correlation between it and a particular language feature (e.g. Pericliev, 2004).

paradigmatic redundancy, Lupyan and Dale (2010) emphasise the importance of the features which increase redundancy. Large numbers of irregularities, for example, would make a language less compressible, not more, and would imply that languages spoken by smaller groups are less redundant and compressible, contra Lupyan and Dale (2010, Text S10, Table S2).

1.6 Research focus and plan for experiments

Non-linguistic explanations for cross-linguistic variation in language features, as discussed above, are largely dependent on correlational studies, or theories with very limited empirical evidence to support them. There is little evidence for proposed explanatory mechanisms, and cases where it is very unclear what such mechanisms would be. There are also controversies, with reanalyses of data contradicting original conclusions, concerns about autocorrelation, and identification of different mechanisms which explain the same result.

My contribution to the field is to test some of these hypotheses, focusing on the proposed mechanisms which could explain observed correlations or more theoretic work, through a series of experiments which consider the effect of language learning and use on the languages themselves. I specifically focus on the claims of sociocultural determination of linguistic complexity, which present us with particularly clear and experimentally testable hypotheses (Nettle, 2012). For the remainder of this thesis, I focus on two in particular: languages spoken by more people being less complex (Lupyan and Dale, 2010), and more esoteric communicative contexts resulting in more complex language (Thurston, 1994; Wray and Grace, 2007; Trudgill, 2011).

As it is unclear how number of speakers could itself affect morphological complexity, an explanatory mechanism needs to be identified (Nettle, 1999a, 2012; Bybee, 2011; Dahl, 2011). Daniel Nettle (2012) summarises 3 potential mechanisms by which larger numbers of speakers could correlate with relative morphological simplicity: (cultural) drift, different degrees of variability in linguistic input data, and the maturational learning differences between adults and children.

Analogous to genetic drift, cultural drift may have a more pronounced effect in smaller populations (Boyd and Richerson, 1985). Nettle (1999a) therefore proposed that linguistic change may be faster in smaller populations, increasing the likelihood of less optimal, and therefore possibly more non-functionally complex, communication systems developing. In Nettle (2012), however, he notes the problems with this proposal, not least the lack of empirical evidence demonstrating that there is actually a faster rate of change in smaller populations. Language change may actually be slower in smaller populations, with smaller groups being more likely to converse linguistic norms (Trudgill, 2004a; Wray and Grace, 2007; Trudgill, 2004a). A more functional explanation (i.e. one relating to the language learning and usage needs and preferences of a group of speakers) would therefore appear more likely.[13]

Chapter 2 describes four experiments which investigate the effect of input variability, and the proposal that the linguistic input of languages spoken by greater numbers of speakers is such that the cross-generational transmission of complex morphology is less likely (Nettle, 2012).

Chapter 3 then describes three experiments which assess the effect of adult learning of complex morphology, considering both how an individual adult learner will acquire a simplified

[13] In spite of his claims that language is "hugely dysfunctional", as discussed in Section 1.3.6, even Gil (2009, p. 32) accepts that there is still "overwhelming evidence showing that diachronic change can be functionally motivated". Languages being more complex than necessary would therefore not preclude languages undergoing some functional adaptation to different sociocultural factors.

version of a target language, and the process by which such individual-level simplifications could lead to changes at the population-level of a language (Kirby, 1999; Lupyan and Dale, 2010; Nettle, 2012).

While there is some consensus about the possibility of contact and adult learning having simplifying effects on language, complexifying mechanisms also need to be identified if we are to explain where the more complex features came from in the first place (Thurston, 1994). In Chapter 4, I consider how esoteric communication may account for this complexity. Four experiments investigate the emergence and development of communicative conventions, assessing the effects of social network size and density, and shared knowledge on linguistic complexity (Thurston, 1994; Wray and Grace, 2007; Trudgill, 2011).

In Chapter 5, I then consider the eleven experiments together, and their implications for the claims that sociocultural factors can determine the degree of a language's complexity.

2 Speaker input variability

2.1 Introduction

In this chapter, I investigate the proposal that different degrees of "heterogeneity in the learning set" (Nettle, 2012, p. 1833), the amount of variability in a learner's linguistic input, could provide an explanation for population size determination of linguistic, specifically morphological, complexity (Lupyan and Dale, 2010). This core of this investigation has been published in PLoS ONE and this paper is included in full in Section 2.2. The two experiments in the paper were inspired by two earlier experiments, and I describe these in Sections 2.3 and 2.4, respectively.[14] A more detailed description of the post-test interviews which followed Experiment 2 (Atkinson et al., 2015, p. 15) is also given in Section 2.5.

2.1.1 A note on linear mixed effects analyses

Each of the experiments, both in this chapter and in Chapters 3 and 4, employ at least one linear mixed effects analysis. To save unnecessary replication, I have included an overview of the approach I take in Appendix A.

2.2 Atkinson, Kirby, and Smith (2015): Experiments 1 and 2

The following 20 pages present the paper in its entirety. The co-authors are my Ph.D. supervisors, Kenny Smith and Simon Kirby.

2.2.1 Author contributions

As indicated at the end of the paper, all three authors contributed to the design of the experiments. I ran the experiments and analysed the data. I also wrote the manuscript, with the other authors commenting on earlier drafts.

[14] The motivations for these two experiments were the same as for those in the paper, and their methodologies are very similar. Therefore though they were completed before Experiments 1 and 2, I have included them afterwards so as not to replicate the published introductory and methodological sections.

PLOS | ONE

RESEARCH ARTICLE

Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages

Mark Atkinson*, Simon Kirby, Kenny Smith

Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

* m.d.atkinson@sms.ed.ac.uk

OPEN ACCESS

Citation: Atkinson M, Kirby S, Smith K (2015) Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. PLoS ONE 10(6): e0129463. doi:10.1371/journal.pone.0129463

Academic Editor: Christine A Caldwell, University of Stirling, UNITED KINGDOM

Received: February 11, 2015

Accepted: May 8, 2015

Published: June 9, 2015

Copyright: © 2015 Atkinson et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Raw data files are available at the Edinburgh DataShare repository of the University of Edinburgh: http://dx.doi.org/10.7488/ds/213.

Funding: MA is supported by an Arts & Humanities Research Council (http://www.ahrc.ac.uk) PhD Studentship (grant number AH/K503010/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

A learner's linguistic input is more variable if it comes from a greater number of speakers. Higher *speaker input variability* has been shown to facilitate the acquisition of phonemic boundaries, since data drawn from multiple speakers provides more information about the distribution of phonemes in a speech community. It has also been proposed that speaker input variability may have a systematic influence on individual-level learning of morphology, which can in turn influence the group-level characteristics of a language. Languages spoken by larger groups of people have less complex morphology than those spoken in smaller communities. While a mechanism by which the number of speakers could have such an effect is yet to be convincingly identified, differences in speaker input variability, which is thought to be larger in larger groups, may provide an explanation. By hindering the acquisition, and hence faithful cross-generational transfer, of complex morphology, higher speaker input variability may result in structural simplification. We assess this claim in two experiments which investigate the effect of such variability on language learning, considering its influence on a learner's ability to segment a continuous speech stream and acquire a morphologically complex miniature language. We ultimately find no evidence to support the proposal that speaker input variability influences language learning and so cannot support the hypothesis that it explains how population size determines the structural properties of language.

Introduction

Languages evolve, adapting to pressures which arise from their learning and use [1]. As these pressures may be different in different physical, demographic and sociocultural environments, non-linguistic factors may systematically determine linguistic features [2–5]. Identifying those factors which specifically affect the structural properties of language, and establishing the mechanisms by which they operate, will shed light on why languages exhibit different degrees of grammatical complexity [4] and how individual-level learning interacts with the sociocultural features of a speech community to result in group-level language features [6–8]. It may also aid our understanding of typological and psycholinguistic constraints on language [2–4], as well as provide clues as to the emergence of structure in the early language of our species [2].

At the level of the individual learner, the language an individual acquires depends on the specific linguistic input they receive, the context in which it is transmitted, and the way that input interacts with the learning abilities and biases of the learner [5, 8, 9]. Across different types of groups in different environments, there may be systematic differences in the input data learners receive and the effect it has on their developing languages. This may explain observable differences in languages spoken by different types of social groups in different environments [2–5, 10].

Here we consider one particular feature of the linguistic input, the degree of homogeneity in the data arising from the number of speakers who provide it. It has been suggested that this difference may have systematic effects on the acquisition of complex morphology, and that this may result in the simplified morphological systems seen in the languages of larger groups [5].

Speaker input variability and phoneme acquisition

Variability in linguistic input can arise at multiple levels of analysis, from different lexical items or word orders being used to convey the same semantic information down to subtle variability in the realisation of phonemes. One source of the latter kind of variability is the differences in the idiosyncratic pronunciations of the speakers who provide the input. This results from dialectal differences and variable speech rates, as well as anatomical differences amongst the speakers, such as the length and shapes of their oral and nasal cavities [11, 12]. *Speaker input variability* may therefore be increased either by the pronunciation being less homogeneous across the speakers, or by the data being provided by a greater number of speakers [5].

A number of studies have demonstrated the effect that input variability can have on the acquisition of phonemic (or tonal [13] contrasts. These studies consider adult second language acquisition and typically focus on Japanese learners of English attempting to acquire the contrast between /l/ and /r/. Input variability is manipulated by either exposing learners to target phonemic distinctions in a greater number of lexical contexts, or by considering the effect of High Variability Phonetic Training (HVPT), where the learner is simply exposed to "natural words from multiple talkers" [14, p. 3267]. Both types of variability aid discrimination of target phonemic contrasts [12, 15–17], with a direct comparison of the two manipulations finding HVPT more effective than context variability [18]. The effects of HVPT have also been confirmed in discrimination tasks involving familiar and novel speakers [15, 16, 18], for retention of phonemic boundaries 6 months after training [15], and in learner productions [16, 19].

This evidence that increasing speaker input variability can aid phoneme acquisition, and by extension minimal pairs of a lexical set, is alone enough to suggest that its effect on other aspects of language acquisition is worth investigation. But it has also been proposed that speaker input variability may explain how non-linguistic features of a speech community could influence structural features of its language.

Sociocultural determination of linguistic structure

A body of work has already aimed to identify the sociocultural factors which influence non-structural features of language. For example, the number, specificity and semantic complexity of lexical items results from a group's need for and ability to maintain distinctions [2, 3, 20]: distinguishing amongst different types of sheep will be more useful to sheep farmers than other social groups, and so the lexicon of a British sheep farmer will include terms such as *gimmer*, *freemartin* and *rigger*, which may well be unfamiliar terms to other speakers [3]. Perhaps more speculatively, phoneme inventories and phonotactic constraints are thought to have adapted to have a greater proportion of more sonorant phonemes in environments which favour more distal communication, such as in warmer climates or where there is less vegetation [21–23]. The size of a language's phoneme inventory may also be influenced by its number of speakers: languages of larger groups have been claimed to have larger phoneme sets [24–26].

There is a growing interest in how demographic or sociocultural factors may determine **structural** features of a language [5]. Wray and Grace [2] discuss how different sizes and types of social group might influence systematic differences in the complexity of their languages, considering two extremes of communication: esoteric, or intra-group, and exoteric, or inter-group, communication. They argue that esoteric communication, as used by speakers in small, unified social contexts where a lot of information can be presupposed, will be more complex. There will be a greater number of irregular and opaque features, a higher degree of morphological complexity with a greater number of irregularities (note that use of morphological strategies over lexical is in itself likely to result in an increase in the number of irregular forms [27]) and more derivational constraints leading to increased suppletion. Conversely, exoteric communication is that employed by larger groups, with a large amount of interaction conducted between strangers and therefore with more limited shared information for interlocutors to rely on. Such communication will be less grammatically complex, characterised by one-to-one relations between form and meaning, allomorphy, regularity, transparency, flexibility of expression and compositionality of signals. Wray and Grace argue that the complex nature of esoteric communication is more representative of the "default" psycholinguistic preference for less regular and transparent language, and so will be the result of languages which prioritise child language learning and the communicative needs of more intimate social groups. Simpler, exoteric, communication is then a "consequence[] of talking to strangers" [2, p. 543], where the language has adapted to the needs of adult language learning. Trudgill [3] also argues that more complex languages are more likely to be found in situations where there is less contact with other languages, higher social stability, smaller speech communities, denser social networks and more "communally-shared information" [3, p. 146].

These claims receive empirical support from work by Lupyan and Dale's study of the correlation between demography and morphological complexity [10]. Following previous work investigating the relationship between the number of speakers of a language and grammatical complexity [28, 29], they investigate 2,236 languages using data from the World Atlas of Language Structures database [30], considering 28 structural features relating to each language's morphological type, case system, verb morphology, agreement, possibility and evidentials, negation, plurality, interrogatives, tense, possession, aspect, mood, articles, demonstratives and pronouns. Controlling for language family and geographic location, they find that languages with larger populations, spoken over larger areas and in contact with a greater number of other languages tend to be characterised by lower morphological complexity and the greater use of lexical strategies to make semantic distinctions. They found that population size had the most predictive power, and specifically claim that languages spoken by a greater number of people have less complex inflectional morphology. More recently, simulations of language learning have also supported the proposal that the languages of larger groups are likely to have a greater number of simpler conventions which are easier for a learner to acquire [31].

Speaker input variability and structural complexity

Discovering a correlation between a non-linguistic factor such as number of speakers and the structural features of a language is not satisfactory in itself: a causal mechanism needs to be identified to explain why and how a proposed determinant could have such an effect. Population size itself may actually not be the most informative predictor. There may instead be a more direct determinant, some aspect of society or environment which is itself correlated with larger groups [5]. Alternatively, the effect may be the result of the interaction of a number of factors [3], with features, such as cultural complexity [32, 33], whether or not the language has a written form [2, 34] and language age [35], also having some influence.

One proposed explanation, discussed by Nettle [5], is the differing degrees of speaker input variability encountered by learners in different sized groups. Nettle suggests that an individual's social network will be more constrained in smaller populations. The input they receive is therefore likely to be more homogeneous, being provided by a smaller number of speakers, or otherwise exhibiting less inter-speaker variability due to the reduced possibilities for dialectal differences. In larger groups, the learner is part of a larger social network, and so the input they receive is likely to be more variable. Nettle proposes that increased variability makes morphological distinctions, which are often based on minimal phonological differences, more difficult to acquire and hence less likely to survive cross-generational transfer. With the loss of these comparatively subtle distinctions, an alternative strategy is necessary if the same semantic distinctions are to be maintained. This is likely to be an innovated, structurally more simple, lexical strategy [5].

A challenge for this proposal is to explain why greater input variability aids phoneme acquisition yet hampers the acquisition of morphology [5]. One solution is to note the very different roles that increased variability may have in each case. In the acquisition of a phoneme set, higher variability provides more information about the group-level distribution of a phoneme and so aids the maintenance of phonemic distinctions. In the acquisition of morphology, however, it may simply increase the noise in the input and make the target less accessible to the learner. Such an account may explain why languages of larger groups appear to have both larger phoneme sets [24–26] (though see [36]) and simpler morphological systems [5, 10].

In the remainder of this paper we describe two experiments designed to test the effects of speaker input variability on language acquisition, and therefore test the plausibility of speaker input variability as a mechanism explaining how group size influences morphological complexity. In Experiment 1, we extended previous work on statistical learning to consider whether the effect of speaker variability in phoneme acquisition can be extended to word segmentation. In Experiment 2, we tested the effect of speaker input variability on the learning of a morphological system. To anticipate our results: we find no evidence that increased speaker input variability impedes (or indeed facilitates) the learning of morphology, therefore throwing some doubt on the viability of this mechanism.

Experiment 1: word segmentation

In their seminal study investigating the abilities of learners to use distributional cues to segment continuous linguistic input, Saffran et al. [37] demonstrated that adults were able to segment words from a speech stream using only the transitional probabilities between consonant-vowel (CV) syllables. These abilities have since been extended to infants [38], natural speech [39], larger learning sets [40], the acquisition of multiple languages [41], non-linguistic auditory tasks [42], equivalent capabilities in the visual field [43, 44] and even to other species [45].

The transitional probability between the elements of an input stream is computed by dividing the frequency of a pair of units *XY* by the frequency of the unit *X*. A higher probability then indicates that the presence of element *X* more strongly predicts the subsequent presence of *Y*. An example, taken from Saffran et al. [37, p. 610], considers the syllable as the unit of analysis and the English word *baby* (/beɪ.bi/). /beɪ/ is a relatively high-frequency syllable, which will be followed by /bi/ some of the time. But it can also be followed by other syllables, both within a word, as in *bacon* or *baker*, or across a word boundary, as in *Bay of* or *obey the*. Since words can be freely combined (within the syntactic constraints of a language), the predictability of a second element in a pair of syllables within words will generally be higher than those which span a word boundary, and so the probability of, for example, /bi/ following /bei/ is likely to be higher than /ðə/ following /bei/. Therefore the transitional probability of /beɪ .bi/ would then be higher than /bei#ðə/. Transitional probabilities therefore form a cue which can be used to identify the components of an input stream: while in the statistical learning literature these are typically glossed as words, the same logic applies to the segmentation of complex signals built by productive morphological processes.

Determining the morpheme boundaries of input data is one of the first steps in the acquisition of a morphological system [37, 40]. Therefore if increased speaker input variability makes the segmentation of a speech stream more difficult, a learner may find the acquisition of complex morphology more challenging; this may eventually result in the language simplifying as it is transmitted from learner to learner [5]. To assess this, we adapted the experimental design of Saffran et al. [37] to investigate whether or not there is an effect of the number of speakers who provide the input. To our knowledge, this is the first investigation of the effect of speaker input variability on word segmentation and the first attempt to see if the findings of the HVPT studies can be extended to other aspects of language acquisition.

Materials and methods

This experiment was approved by the Linguistics and English Language Ethics Committee of the University of Edinburgh. Written consent was provided by all participants before taking part.

The methodology for this experiment was based on the first experiment described in Saffran et al. [37], with an additional manipulation of speaker input variability. We assessed the ability of adult learners to discriminate between words and non-words in forced-choice testing after exposure to a continuous speech stream. In the single speaker condition, the learner's input came from a single speaker; in the multiple speaker condition, the input was instead spread among 3 different speakers.

Following Saffran et al. [37], four consonants (p, t, b, d) and three vowels (a, i, u) were used to construct an inventory of 12 CV syllables, from which six trisyllabic words were created (*babupu*, *bupada*, *dutaba*, *patubi*, *pidabu*, *tutibu*). An aural stimulus was constructed by concatenating the words of the language into a continuous speech stream, lacking acoustic cues to word boundaries. 300 tokens of each word were randomly ordered, with words then eliminated so that no adjacent words were the same. In contrast to Saffran et al. [37], and to reduce any influence of the order of a particular input string, we generated 24 such input strings, each independently randomised, and used each once only in each experimental condition. In each string, the transitional probabilities within a word were greater than the transitional probabilities across a word boundary, as in the original study. For each of the 24 input strings, 6 trisyllablic non-word foils were randomly constructed using the 12 syllables of the CV inventory, but with the stipulation that the transitional probabilities between the syllables within the speech stream was 0. One foil set, for example, was *bubidi*, *tabidi*, *tatupa*, *dubati*, *bitapi* and *tupati*.

As in previous studies [40, 46], the target words, input streams and foils were created using the MBROLA speech synthesis package [47], with a CV syllable duration of 278ms [37], of which 60ms was assigned to the consonant. 4 diphone databases were used to construct each target, input stream and foil for each of 4 different speakers. A constant F0 of 100Hz was assigned to 3 male voices (en1, us2, and de1) [40, 46], and 200Hz for 1 female voice (us1) [48]. Use of synthesised speech ensured that there were no acoustic cues to word boundaries.

Participants

48 native English speakers (10 male; aged between 18 and 33, mean 21.1) were recruited using the Student and Graduate Employment (SAGE) database of the Careers Service of the University of Edinburgh. Each was compensated £5.50.

Procedure

As in Saffran et al. [37], the participants were told they were going to listen to a "nonsense" language, which contained words, but no meanings or grammar. They were told "Your task is to try and figure out where the words begin and end. You don't know how many words there are, nor how long they might be". To justify the unnaturalness of the monotone stimuli, the language was described as a "robot" language, with the speakers being native robot speakers of the language. Though explicit instruction may influence learning [49–52] (though see [53]), it was not anticipated that replicating the previous study's instructions [37] would negatively affect the participants' ability to identify the word boundaries.

Following Saffran et al. [37], the training strings were split into 3 blocks of approximately 7 minutes each, presented with a 5 minute rest after the first and second blocks. In the single-speaker condition (24 participants), a participant was trained using a single voice, with the voice used counterbalanced across participants (6 participants being trained by each of the 4 voices). In the multiple-speaker condition (24 participants), a participant was trained using 3 of the 4 different voices, with the voices used counterbalanced across participants (6 participants being trained by each of the 4 possible combinations of 3 voices). In this multiple speaker condition, each of the training voices provided a third of the input in each of the 3 blocks in a random order. The multiple-voice audio files were created using Audacity 2.0.5, with 5 seconds of cross-fade between speakers, so as not to provide any additional cues as to the word boundaries at the changeover points. The difference between the training regimes in each condition is illustrated in Fig 1.

Training was followed by two forced-choice testing blocks: one with the stimuli presented by the speaker(s) used in training and one using a novel speaker. In each test block, a participant was presented with all 36 possible word-foil pairings, presented in a random order. For each pairing, the word and foil were presented in a random order with 500ms of silence between them. The participant was required to "decide which of the words is from the robot language". There was then a 2 second pause before the next pairing.

The familiar-voice block was designed to replicate Saffran et al. [37], while the novel-voice test was included to investigate any possible effect of multiple-speaker training and the comprehension of an unfamiliar speaker, following similar findings in HVPT [15, 16, 18]. To

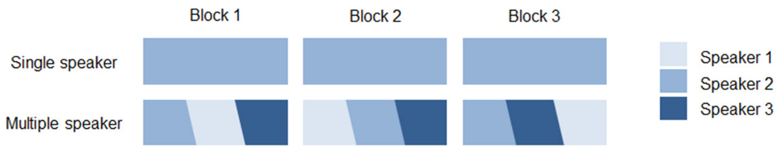


Fig 1. Example training regimes for participants in the single and multiple speaker conditions.

doi:10.1371/journal.pone.0129463.g001

control for any ordering effects, the blocks were counterbalanced so that half the participants in each condition were presented with the familiar speaker test first and half with the novel speaker test first.

For a participant in the single-speaker condition, the familiar-voice testing block used the same voice as in training. The novel-voice block used one of the other 3 voices. Over the set of single-speaker participants, each combination of familiar voice and novel voice was used twice. For a participant in the multiple-speaker condition, each of the voices from the training were used for a third of the testing pairings in the familiar-voice block. The novel-voice block then used the only voice not used in training.

The experiment was written and run in Matlab (R2013b) with the Psychtoolbox extensions.

Analysis and results

Learning was assessed by counting the number of times the word was correctly identified in the word-foil test pairings. The maximum score in each block was 36, with chance performance 18. The results are shown in Fig 2.

We performed a linear mixed effects analysis using R [54] and *lme4* [55]. We fit a maximal model [56] with logit regression including *condition* (single speaker or multiple speaker), *speaker identity* (familiar or novel), *order of tests* (familiar speaker test first or second) and the interaction of *condition* and *speaker identity* as (centred) fixed effects, with *participant identity* as a random effect. The interaction of *condition* and *speaker identity* was included to see if there was any effect of participants in the multiple speaker condition being better at distinguishing words from foils when listening to unfamiliar speakers, following similar findings in HVPT [15, 16, 18]. The model was significantly better than the equivalent null model ($\chi^2(4)$ = 52.457, p <0.001). The intercept was significantly different from zero (β = 0.343, SE = 0.065, p <0.001), reflecting that, averaging across all our data, participants performed significantly

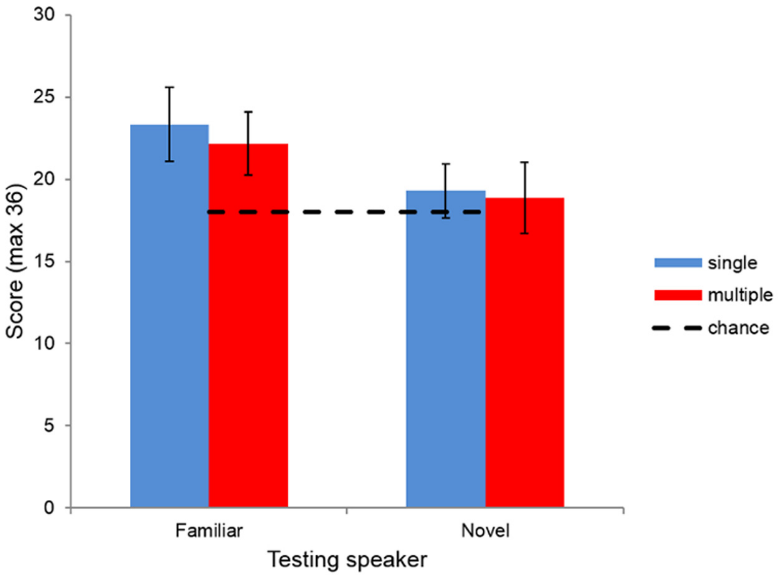


Fig 2. Average scores for each condition in both familiar speaker and novel speaker testing blocks. Error bars represent 95% confidence intervals of the mean in each case.

doi:10.1371/journal.pone.0129463.g002

better than chance (participants were 1.41 times as likely to produce a correct response on test as incorrect, corresponding to an accuracy of 58%). There were significant contributions of *speaker identity* (β = -0.439, SE = 0.070, p <0.001) and *order of tests* (β = -0.258, SE = 0.070, p <0.001). There were no effects of *condition* (β = 0.097, SE = 0.130, p = 0.457) or the interaction of *condition* and *speaker identity* (β = -0.093, SE = 0.141, p = 0.508).

This analysis suggests that the participants were able to use the distributional cues in their training strings to discriminate between words and non-words, replicating the result of Saffran et al. [37]. Performance was better in the familiar voice testing: participants were 1.76 times as likely to produce a correct response as incorrect, corresponding to an accuracy of 64%; in the novel voice testing, they were 1.13 times as likely, corresponding to an accuracy of 53%. Greater performance in the familiar voice testing supports the HVPT findings that distinguishing words is easier when they are presented by familiar speakers [15, 16, 18]. Scores in the second test blocks were also on average lower than those in the first, suggesting either an effect of participant fatigue or interference from the first block. There is also evidence of the participants being able to generalize their training input to a novel speaker. Considering only the novel voice testing data presented in the first block, a linear mixed model with logit regression and no fixed effects and *participant identity* as a random effect had an intercept significantly greater than zero (β = 0.184, SE = 0.090, p = 0.041): participants were 1.20 times as likely to produce a correct response as incorrect, corresponding to an accuracy of 55%.

Conclusions of Experiment 1

The lack of a difference between the conditions extends Saffran et al.'s [37] result to the case where the training data is presented by multiple voices, suggesting that segmentation of continuous speech may not be affected by the number of speakers who provide it. We have no evidence, however, that the effects of speaker input variability on phonemic acquisition can be extended to a learner's ability to segment their linguistic input. Though the acquisition of morphology involves much more than segmenting input, determining word boundaries is still a crucial part of this process [57]. Therefore there is no evidence to support the proposal that speaker input variability could influence morphology learning.

Experiment 2: learning morphology

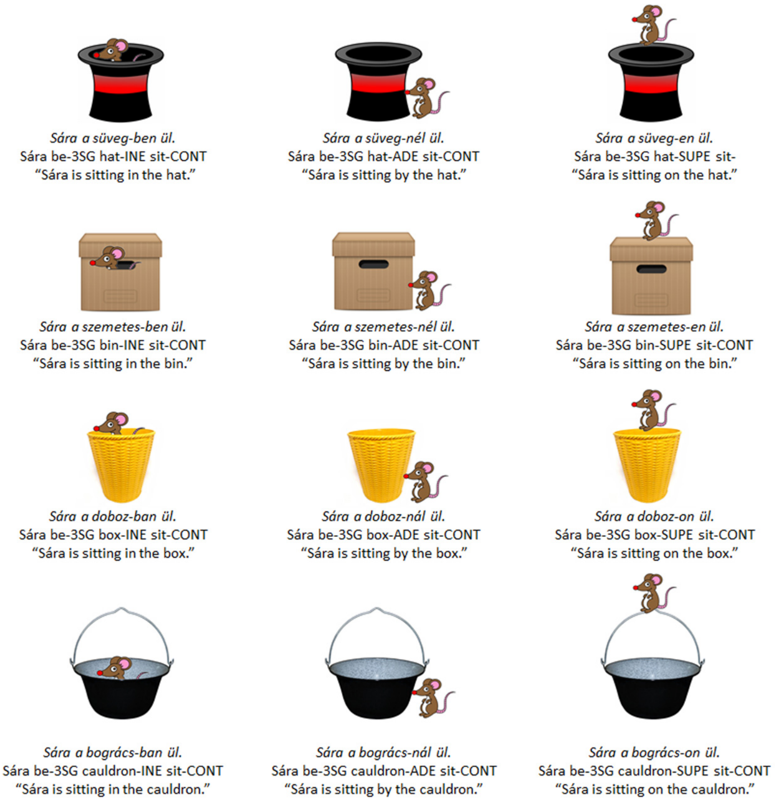
Our first experiment, in assessing the effect of speaker input variability on the ability of a learner to isolate and identify individual morphemes in a speech stream, investigated a crucial part of an individual's acquisition of a morphological system [37, 40]. But the learner has to do more than distinguish morpheme boundaries: they also have to relate the isolated components to meanings, be able to recombine them to create grammatically permissible utterances which convey particular semantic information, and then be able to produce these utterances. We conducted a second experiment which more closely reflects the full range of processes involved in morphology learning and so more thoroughly tests the effect of speaker input variability on the acquisition of morphology, assessing learner abilities to orally acquire a morphologically-complex miniature language.

Materials and methods

This experiment was approved by the Linguistics and English Language Ethics Committee of the University of Edinburgh. Written consent was provided by all participants before taking part.

We asked participants to learn a miniature language based on 12 sentences of Hungarian. Hungarian has an extensive nominal case system in which nouns are (barring rare exceptions) obligatorily marked with case-indicating suffixes [58–60]. The particular form of a suffix is also often dependent on vowel harmony, with a [+back] feature in the initial vowel of the noun stem spreading throughout the stem and its suffixes [60–62]. Hungarian has 14 vowels, including a phonemic contrast between long and short vowels. The 6 [+back] vowels (with corresponding International Phonetic Alphabet representation) for the purposes of vowel harmony, are *a* (/ɔ/), *á* (/a:/), *o* (/o/), *ó* (/o:/), *u* (/u/) and *ú* (/u:/) [59]. For example, the inessive form of *város* /va:roʃ/, "city", is *városban* /va:roʃbɔn/, "in the city", while the corresponding form of *szék* /se:k/, "chair", is *székben* /se:kbɛn/, "in the chair" [58]. In the first case, the [+back] feature of *á* /a:/ spreads through the suffix, which takes the back vowel of *a* /ɔ/ in *-ban*, while in the second, the [-back] feature of *é* /e:/ results in the alternation *-ben* with the front vowel /ɛ/.

Our target language used three cases: the inessive ("in"), adessive ("by" or "at") and superessive ("on"). These were selected as they each require different affix variants dependent on the initial vowel in the noun stem [58] and were semantically easy to represent using simple and static visual stimuli. 12 images were created in which a cartoon mouse was shown located either in, next to, or on top of one of four containers: a hat, a wastepaper bin, a box and a cauldron. Two of the containers, *süveg* /ʃyvɛg/ ("hat") and *szemetes* /sɛmɛtɛʃ/ ("bin"), have [-back] initial vowels, while the other two, *doboz* /doboz/ ("box") and *bogrács* /bogra:tʃ/ ("cauldron"), have [+back]. The target language therefore includes semantically-redundant alternations within the case-marking affixes. Hungarian sentences describing each of the images then comprised the target language. The complete set of images and labels is given in Fig 3.



Sára a süveg-ben ül.
Sára be-3SG hat-INE sit-CONT
"Sára is sitting in the hat."

Sára a süveg-nél ül.
Sára be-3SG hat-ADE sit-CONT
"Sára is sitting by the hat."

Sára a süveg-en ül.
Sára be-3SG hat-SUPE sit-
"Sára is sitting on the hat."

Sára a szemetes-ben ül.
Sára be-3SG bin-INE sit-CONT
"Sára is sitting in the bin."

Sára a szemetes-nél ül.
Sára be-3SG bin-ADE sit-CONT
"Sára is sitting by the bin."

Sára a szemetes-en ül.
Sára be-3SG bin-SUPE sit-CONT
"Sára is sitting on the bin."

Sára a doboz-ban ül.
Sára be-3SG box-INE sit-CONT
"Sára is sitting in the box."

Sára a doboz-nál ül.
Sára be-3SG box-ADE sit-CONT
"Sára is sitting by the box."

Sára a doboz-on ül.
Sára be-3SG box-SUPE sit-CONT
"Sára is sitting on the box."

Sára a bogrács-ban ül.
Sára be-3SG cauldron-INE sit-CONT
"Sára is sitting in the cauldron."

Sára a bogrács-nál ül.
Sára be-3SG cauldron-ADE sit-CONT
"Sára is sitting by the cauldron."

Sára a bogrács-on ül.
Sára be-3SG cauldron-SUPE sit-CONT
"Sára is sitting on the cauldron."

Fig 3. Complete target language with corresponding images.

doi:10.1371/journal.pone.0129463.g003

Three native speakers of Hungarian (1 female) were recruited to construct the aural training data. In an attempt to have as natural-sounding a stimuli set as possible, they were recorded producing each sentence three times, with the second production used in the experiment.

Participants

40 participants (16 male; aged between 18 and 42, mean 21.4) were recruited using the Student and Graduate Employment (SAGE) database of the Careers Service of the University of Edinburgh, with non-native speakers of English and current and former students of linguistics excluded. Participants were asked to list the languages they could speak or understand, indicating their proficiency in each case. No applicants reported any prior knowledge of Hungarian or any other Uralic language. Participants were required to attend 3 sessions of approximately 20 minutes on consecutive days and at the same time each day. Each was compensated £12 on completion. Data for one further participant was rejected as they did not attend after the first session, and another participant was recruited in their place.

Procedure

Each participant took part in 6 rounds of training and testing, 2 on each day. For each participant, 8 of the 12 target language sentences formed the training data, which were randomly selected with the constraints that two sentences described each container, that each case was represented at least twice and each alternation was represented at least once. The training data was therefore sufficient (in principle) to reconstruct the entire target language, including the 4 unseen sentences.

20 participants were randomly assigned to the single-speaker condition, where the 8 training sentences were produced by the same, randomly-selected speaker throughout the experiment. Each of the 3 speakers was assigned to at least 6 participants. In the multiple-speaker condition, the 8 training sentences were randomly assigned to the 3 speakers with the constraint that at least 2 sentences were presented by each speaker. Each training sentence was then presented by the same speaker throughout the experiment.

In each training round, the learner was exposed to 5 independently randomly sorted passes of the entire training set of 8 image-label pairings. For each item, the participant was first shown the image for 2 seconds in silence, before being played the appropriate audio file and then given 6 seconds to attempt to repeat what they had heard. Advance to the next item was automatic. Before the initial training stage, the learner was given two additional randomly-selected training items to check their comprehension of the task.

Each training stage was followed immediately by a test. The learner was required to orally label the entire set of 12 images (both the 8 seen in training and the 4 novel), presented in a random order. Once an image had been displayed for at least 3 seconds and the participant had had the opportunity to produce a label, any key press on the keyboard advanced the test to the next item.

The experiment was written and run in Matlab (R2010a) with the Psychtoolbox extensions. Audio data was collected using the ProTools LE software and the Digidesign 003 audio interface.

Analysis and results

Production of the noun stems

For each participant utterance, the noun stem and case-marking suffix were segmented and transcribed using the following phoneme set: /y, ɛ, a, ɔ, ə, m, n, ŋ, b, p, d, t, g, k, f, v, s, ʃ, z, ʒ,

Table 1. Articulatory feature values for vowels.

Vowels	/y/	/ɛ/	/a/	/ɔ/	/u/	/ə/
Height	1	0.5	0	0.5	1	0.5
Forwardness	1	1	0.5	0	0	0.5

doi:10.1371/journal.pone.0129463.t001

t͡ʃ, d͡ʒ, w, l, r, j/. Due to hesitations and pauses in the productions, it was not possible to transcribe meaningful length distinctions. Production of the noun stems was then assessed by considering a modified normalised weighted Levenshtein edit distance between the produced stem and target, with distance from individual phonemes based on the articulatory feature values provided by Connolly [63]. Feature values for the vowels and consonants of our transcription set are given in Tables 1 and 2, respectively. We have assumed that all unvoiced plosives are aspirated, have set the sulcral values for /ʒ/, /t͡ʃ/, /d͡ʒ/, /w/, /l/, /r/ and /j/ ourselves, and have taken average values for double articulators.

Following the recommendations of previous work [64, 65], insertions and deletions were given an edit cost of 1, and replacement of a vowel with a vowel or a consonant with a consonant a maximum value of 0.8. Replacing a vowel with a consonant or vice versa incurred a cost of 1. The distance between two phonemes was calculated by taking the sum of the absolute values between each of their features. So, for example, the distance between /y/ and /a/ is calculated by $|1 - 0| + |1 - 0.5| = 1.5$, and the distance between /n/ and /t͡ʃ/ by $|0 - 0.5| + |0.85 - 0.85| + |0.85 - 0.85| + |1 - 0.95| + |1 - 0| + |0 - 0| + |0 - 0.8| + |0 - 1| = 3.35$. These distances are then normalised by dividing by the maximum distance within the set of vowels (1.5) or consonants (4.25), and then multiplying by the maximum within-category phoneme replacement factor of 0.8 [64]. A final distance between two strings was then normalised by the length of the longer string, and an accuracy score calculated as 1 minus this value.

For example, consider the distance between the two strings /kam/ and /fi/. Replacing /k/ with /f/ incurs a cost of $(1.3/4.25) \times 0.8$. Replacing /a/ with /i/ incurs a cost of $(1.5/1.5) \times 0.8$ (note that this is the maximum distance between two vowels). Inserting /m/ incurs a cost of 1. Normalising the sum by dividing by the maximum string length of 3, we have a distance measure of 0.682, and so an accuracy score of 1–0.682 = 0.328.

Mean stem accuracy for each of the conditions over the 6 rounds is illustrated in Fig 4.

Table 2. Articulatory feature values for consonants.

Consonants	n	m	ŋ	b	p	d	t	g	k	f	v	s	ʃ	z	ʒ	t͡ʃ	d͡ʒ	w	l	r	j
Aspiration	0	0	0	0	1	0	1	0	1	0.5	0	0.5	0.5	0	0	0.5	0	0	0	0	0
Place	0.85	1	0.6	1	1	0.85	0.85	0.6	0.6	0.9	0.9	0.85	0.8	0.85	0.8	0.85	0.85	0.8	0.85	0.8	0.7
Constrictor	0.85	1	0.6	1	1	0.85	0.85	0.6	0.6	1	1	0.85	0.85	0.85	0.85	0.85	0.85	0.8	0.85	0.85	0.6
Stop	1	1	1	1	1	1	1	1	1	0.9	0.9	0.9	0.9	0.9	0.9	0.95	0.95	0	0	0	0
Nasal	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lateral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Sulcal	0	0	0	0	0	0	0	0	0	0	0	1	0.8	1	0.8	0.8	0.8	0	0	0	0
Double	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0

doi:10.1371/journal.pone.0129463.t002

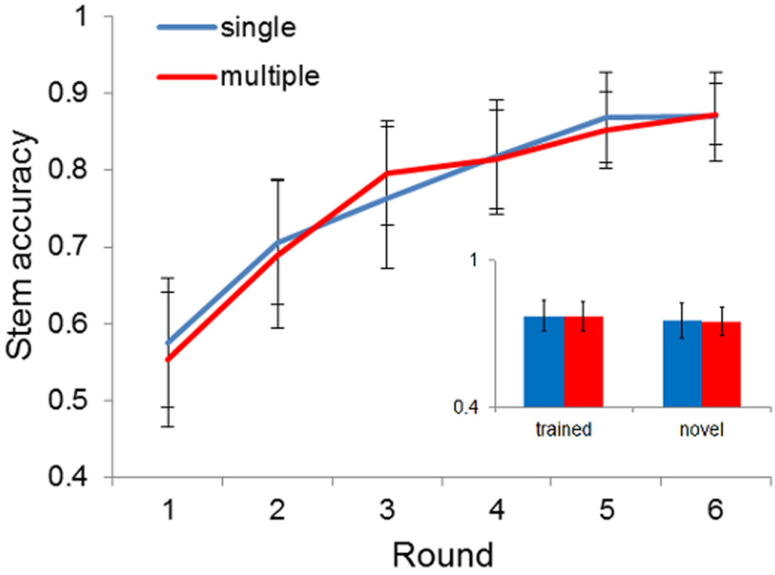


Fig 4. Accuracy of participant productions of target stems. Main graph shows the production scores of the complete target language (the entire set of 12 items). The insert illustrates the minimal difference between the average scores (over all rounds) for the trained and novel items. Error bars represent 95% confidence intervals of the mean.

doi:10.1371/journal.pone.0129463.g004

We performed a linear mixed effects analysis using R [54] and *lme4* [55]. A maximal model [56] included *condition* (single speaker or multiple speaker), *novelty* (whether the target stimulus had been seen in training or not) and *round* and their interactions as (centred) fixed effects. *Participant identity* was investigated as a random effect. This model was significantly better than the equivalent null model ($\chi^2(7) = 628.91$, p <0.001). P-values were estimated from the resultant t-statistics with 2873 degrees of freedom, the number of observations minus the number of fixed parameters in the model [66]. There were significant effects of *round* ($\beta = 0.059$, SE = 0.002, t (2873) = 26.35, p <0.001) and *novelty* ($\beta = 0.020$, SE = 0.008, t (2873) = 2.40, p = 0.016), but no effect of *condition* ($\beta = 0.004$, SE = 0.041, t (2873) = 0.09, p = 0.928) or any of the interaction terms.

This analysis suggests that participant production of the noun stems improved with increased training and testing, and that participants more accurately produced the stems for images they saw in training. No effect of condition suggests that speaker input variability had no effect on acquisition. There is therefore no evidence that the number of speakers who provide the input affects language acquisition in general, and we turn our attention to assessing the claim that it may have a specific effect on morphology.

Production of the affixes

To assess participant acquisition of the morphological system, each produced affix was binary coded using three increasingly stringent measures:

1. Case identification—"1" if and only if the affix unambiguously identified the correct case of the target.

2. Case accuracy—"1" if and only if the affix was an accurate reproduction of one of the alternations for the case of the target.

3. Alternation accuracy—"1" if and only if the affix was an accurate reproduction of the correct, vowel-harmony dependent, alternation of the target.

For example, consider the target suffix for a [-back] stem marking the inessive case, *-ben* /-bɛn/. A production of /-bɛm/ would be coded 1 for case identification, 0 for case accuracy and 0 for alternation accuracy, as while the target case can be unambiguously recovered from the production, the realisation does not exactly match either suffix (corresponding to either [-back] or [+back] stems) which marks the inessive case in the target language. A production of /-bɔn/ would be coded 1 for case identification and 1 for case accuracy, as although the alternation is not appropriate for a [-back] stem, the participant accurately produced one of the suffixes of the target cases, but violated vowel harmony. Only a production of /-bɛn/ would score 1 for all three measures.

The coding for each of the measures was carried out twice. The measurements were first hand-coded directly from the recordings of the participants' productions. These were then compared to calculations of modified normalised weighted Levenshtein edit distances between the transcriptions of the produced affixes and the affixes of the target language calculated using the same methods as described for the stems above. For the case identification measure, we calculated the edit distances between the transcription and each of the 6 suffixes of the whole target language. We then checked that a score of 1 had been coded if and only if the lowest of these edit distance corresponded to the distance between the transcription and one of the two suffixes of the target case. For example, if the target was in the inessive case, we confirmed that a score of 1 was awarded if and only if the edit distance between the transcription and /-bɛn/ or the edit distance between the transcription and /-bɔn/ was lower than all the other distances between the transcription and the other suffixes of the language. For the case accuracy measure, we checked that a hand-coded score of 1 corresponded to the edit distance between the transcription and one of the two suffixes of the target case being 0. For the alternation accuracy measure, we checked that a hand-coded score of 1 corresponded to the edit distance between the transcription and the target suffix being 0.

The results by condition for each measure are shown in Fig 5. Average scores for the whole language are given, along with a comparison of the scores relating to the trained and the novel images.

We performed linear mixed analyses for each measure, using logit regression and maximal models [56] which again included *condition*, *novelty* and *round* and their interactions as (centred) fixed effects. *Participant identity* was again included as a random effect. For all three measures, the fitted model was better than the corresponding null model (Case identification: $\chi^2(7) = 434.12$, $p < 0.001$; Case accuracy: $\chi^2(7) = 218.17$, $p < 0.001$; Alternation accuracy: $\chi^2(7) = 216.71$, $p < 0.001$).

For the case identification measure, there were significant effects of *novelty* ($\beta = 1.112$, SE = 0.100, $p < 0.001$), *round* ($\beta = 0.464$, SE = 0.029, $p < 0.001$) and the interaction of *novelty* and *round* ($\beta = 0.203$, SE = 0.059, $p < 0.001$). There was no effect of *condition* ($\beta = 0.390$, SE = 0.482, $p = 0.419$) or any of the other interaction terms ($p \geq 0.166$):

For the case accuracy measure, there were significant effects of *novelty* ($\beta = 0.936$, SE = 0.110, $p < 0.001$) and *round* ($\beta = 0.321$, SE = 0.029, $p < 0.001$), and an approaching significance effect of the interaction of *novelty* and *round* ($\beta = 0.123$, SE = 0.064, $p = 0.055$). There was no significant effect of *condition* ($\beta = 0.354$, SE = 0.465, $p = 0.447$), or any of the other interaction terms ($p \geq 0.327$).

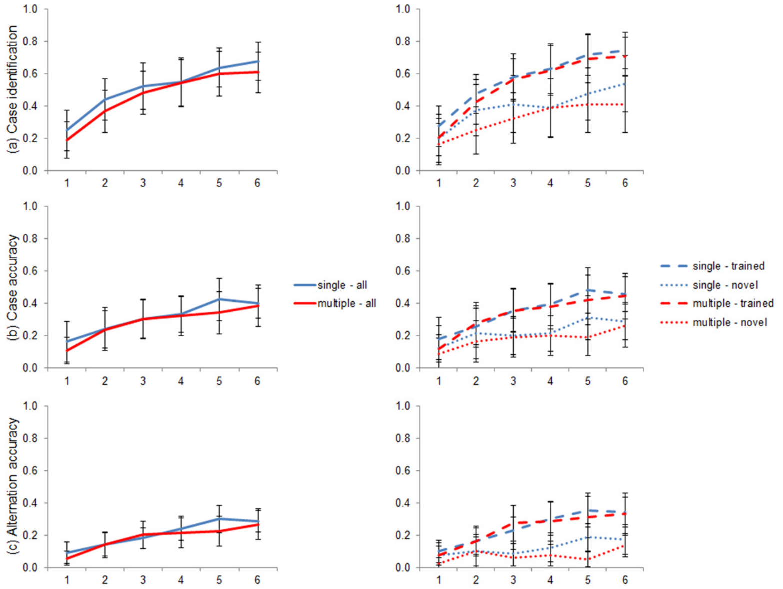


Fig 5. Acquisition of the suffixes. Mean scores by condition are shown for each of the 3 measures, both for the entire target language set (left), and split by training and novel image labels. Error bars show 95% confidence intervals of the mean.

doi:10.1371/journal.pone.0129463.g005

For the alternation accuracy measure, there were significant effects of *novelty* ($\beta = 1.223$, SE = 0.134, p <0.001) and *round* ($\beta = 0.300$, SE = 0.033, p <0.001). There was no significant effect of *condition* ($\beta = 0.468$, SE = 0.389, p = 0.226) or any of the interaction terms (p $\geq$ 0.276).

Conclusions of Experiment 2

Whichever measure we consider, this analysis indicates that participant affix productions improved with increased training and testing, and that the labelling of novel images was worse than that of those seen in training. As in Experiment 1, we find no evidence to support a hypothesis that speaker input variability aids language acquisition, and so again have no support for the suggestion that it should be considered a mechanism by which group size can determine a language's morphological complexity.

Discussion

These experiments provide no evidence to support the hypothesis that speaker input variability may influence language learning beyond the acquisition of phonemic [15, 16, 18, 19] or tonal [13] distinctions. We cannot, of course, rule out the possibility that such variability does affect the acquisition of a morphological system, but that we have failed to capture it. The contrast between our conditions may have been too slight, our samples sizes too small, or our assessment measures too crude. Our experiments may also lack sufficient ecological validity. For obvious reasons of practicality and control, we have attempted to investigate natural language-learning process using adult participants in an artificial laboratory setting. This constitutes an important caveat on our interpretation of our results, particularly in light of some evidence that children may respond to input variability differently to adults [67].

To address such concerns, these experiments could be adapted and extended in a number of ways. The contrast between conditions could be increased simply by having a greater number of speakers in the multiple speaker conditions (for comparison, higher variability in HVPT studies is typically represented by 5 speakers [15, 16, 18, 19]), or the homogeneity of the input could also have been decreased in the multiple-speaker conditions in other ways. Speech-rate differences could have been included in Experiment 1, for example, or a language with a greater amount of inter-speaker variation in pronunciation could have been used to construct the target language in Experiment 2 (Hungarian being notably uniform across its dialects [58]). If the proposed effect is relatively subtle, our experiments may also be improved by larger sample sizes, increased training and testing, or by studying the acquisition of a much larger target language ([46] illustrates how Experiment 1 could be made more challenging by increasing the number and length of target items and reducing training). Frank et al.'s study [40], for example, could be adapted to include a multiple speaker condition. As demonstrated by Saffran et al. [38], adapting Experiment 1 in particular to study the effects in infants children would also be a possibility.

While we would welcome future experimental work in this area, the results of these two experiments do suggest that the speaker input variability effect of phoneme and toneme acquisition cannot (transparently at least) be extended beyond the findings of the HVPT studies, and that it is therefore unlikely to be an explanatory mechanism for how group size determines a language's morphological complexity. We have the same null result in two different experiments, which consider two different stages of the language acquisition process, involve both artificial and natural language learning, and test word segmentation in reception and morphological generalisation in production. Our replication of previous results [37] in the familiar voice test of Experiment 1 in both conditions also suggests that our experimental design and procedure were appropriate, that the participants interpreted the task as intended, and therefore that the result of the second condition is valid. There is also no indication that participants misunderstood the task or adopted particularly obscure strategies in Experiment 2. In a post-experiment interview, 39 of the 40 participants reported their attempts to parse the training sentences to determine which segment corresponded to the container and which to the position of the mouse in the images (the remaining participant said that they would have followed this approach if they had believed that they would have been able to do so successfully in the time available). No participant reported not being able to detect a difference between the training sentences.

If speaker input variability does not affect an individual's learning of morphology, then where does this leave the proposal that input variability could explain how group size determines a language's morphological complexity? One possibility is that increased speaker input variability only limits the cross-generational transfer of morphology when "morphological distinctions rely on a single segment or even sub-segmental phonological change", which is often the case in natural languages [5, p. 1833]. Acquisition difficulties would then arise from learners not being able to detect a difference between minimally different input strings (which was not an issue for the learners in our second experiment). This would suggest, however, that speaker input variability could only be a partial explanation of why languages spoken by more people are simpler. Another possibility is that some type of input variability does have an effect on cross-generational transfer of morphology, but not that which arises at the level of phoneme realisation. Syntactic or lexical variability, for example, may be higher in larger groups and result in simplification across generations of transmission. The predictability of such variability and how it is distributed across speakers would then probably be important factors in determining its effects [68, 69], as would the age of learners who receive it [67]. This is certainly worth further investigation.

It is also worth commentating that even if any effects of input variability (in any form) on language learning can be demonstrated, accounting for how such individual effects can result in language-level change is not necessarily trivial [6], while a convincing demonstration of how and why input variability in larger groups is actually greater is also necessary. We accept that the presumption that an individual's social network is likely to be larger in a larger group is reasonable. However, this may not impact on the variability of the input *which is relevant to language acquisition*, given the influence of other sociocultural factors, such as family size and the role of each parent in childcare [70].

Given these issues and the null results of the experiments, it is worth considering other explanations for how group size could influence morphological complexity. Two other candidate mechanisms are discussed by Nettle [5].

One possibility is that (cultural) drift, which has a more pronounced effect in smaller populations [71], may cause faster rates of linguistic change which result in groups adopting "suboptimal" communicative strategies, such as more complex, overspecified, morphological systems [36, 72]. There are a number of problems with such an explanation, however, not least empirical evidence suggesting that linguistic change may actually be slower in smaller populations [5].

An alternative considers the effect non-native learners can have on a language. Languages spoken by a greater number of people appear to have a greater number of non-native speakers [10]. Older learners are also thought to find the acquisition of complex morphology more challenging compared to other means of encoding the same semantic information. More widely spoken languages might therefore be under similar pressures as those in language contact situations [36]. They will simplify grammatically as they adapt to the needs and preferences of their non-native speakers: "difficult" language features will be filtered out, and more transparent, lexical strategies will be favoured over morphological ones [2, 10, 36, 73, 74]. This in turn leads to a greater reliance on extralinguistic, pragmatic, information, which is again better suited to adult learners [10, 75–77].

A challenge for this account, however, is the focus on simplification of languages due to adult learning: arguably it must also account for the relative complexity of languages with fewer non-native learners [5]. One proposal is that the complex(ified) nature of smaller languages reflects some "default" psycholinguistic state of its speakers, which will be reverted to in the absence of pressures resulting from more exoteric communication [2]. Alternatively, if pressures for language simplification are relaxed, more complex, morphological, strategies may be favoured over syntactic ones in the interests of conciseness and efficiency [3, 5]. Another suggestion is that added complexity in the form of grammatical redundancy may actually aid child language acquisition [10]. It may compensate for the difficulties children have in using pragmatic inference to resolve ambiguous utterances [75–77], or by providing more evidence as to how the signal should be segmented [5, 10]. Further work would be necessary to support such claims [5].

Conclusion

The two experiments described here offer no support for the proposal that speaker input variability can affect the acquisition of morphology. In our first experiment, assessing the ability of adult learners to segment continuous input streams using only the transitional probabilities between syllables, participants were able to discriminate between the words of the training data and foils regardless of whether the input was provided by a single speaker or three. This extends previous work assessing the ability of learners to use distribution cues to parse input data [37] to a case where the input is provided by multiple speakers.

The second experiment, which assessed the acquisition of a miniature language with case-marking affixes, also found no affect of speaker input variability. Therefore we have no evidence to support the proposal that such variability may be a causal explanation for the link between group size and morphological complexity [5, 10]. Given these experimental results, and doubts about the proposed relationship between population size and input variability, we ultimately suggest that it is probably not. We would of course still welcome further tests of speaker input variability's effects, although do believe that investigation of alternative explanations for proposed sociocultural determination of linguistic complexity would be more fruitful.

Acknowledgments

We thank Márton Sóskuthy, Sándor Czettner, and Olga Feher for their help with designing and recording the target language for Experiment 2, and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

Author Contributions

Conceived and designed the experiments: MA SK KS. Performed the experiments: MA. Analyzed the data: MA. Wrote the paper: MA.

References

1. Christiansen MH, Chater N. Language as shaped by the brain. Behav Brain Sci. 2008; 31: 489–508; discussion 509–558. PMID: 18826669
2. Wray A, Grace GW. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. Lingua. 2007; 117: 543–578. doi: 10.1016/j.lingua.2005.05.005
3. Trudgill P. Sociolinguistic Typology: Social Determinants of Linguistic Complexity. Oxford: Oxford University Press; 2011.
4. Dale R, Lupyan G. Understanding the Origins of Morphological Diversity: the Linguistic Niche Hypothesis. Adv Complex Syst. 2012; 15: 1150017. doi: 10.1142/S0219525911500172
5. Nettle D. Social scale and structural complexity in human languages. Philos Trans R Soc Lond B Biol Sci. 2012; 367: 1829–1836. doi: 10.1098/rstb.2011.0216 PMID: 22641821
6. Kirby S. Function Selection and Innateness: the Emergence of Language Universals. Oxford: Oxford University Press; 1999.
7. Smith K, Kirby S. Cultural evolution: implications for understanding the human language faculty and its evolution. Philos Trans R Soc Lond B Biol Sci. 2008; 363: 3591–3603. doi: 10.1098/rstb.2008.0145 PMID: 18801718
8. Smith K. Evolutionary perspectives on statistical learning. In: Rebuschat P, Williams JN, editors. Statistical Learning and Language Acquisition. Berlin: De Gruyter Mouton; 2012. pp. 409–432.
9. Evans N, Levinson SC. The myth of language universals: language diversity and its importance for cognitive science. Behav Brain Sci. 2009; 32: 429–48; discussion 448–492. doi: 10.1017/S0140525X0999094X PMID: 19857320
10. Lupyan G, Dale R. Language structure is partly determined by social structure. PLoS One. 2010; 5: e8559. doi: 10.1371/journal.pone.0008559 PMID: 20098492
11. Mullennix JW, Pisoni DB, Martin CS. Some effects of talker variability on spoken word recognition. J Acoust Soc Am. 1989; 85: 365–378. doi: 10.1121/1.397688 PMID: 2921419
12. Logan JS, Lively SE, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/: a first report. J Acoust Soc Am. 1991; 89: 874–886. doi: 10.1121/1.1894649 PMID: 2016438
13. Wang Y, Spence MM, Jongman A, Sereno JA. Training American listeners to perceive Mandarin tones. J Acoust Soc Am. 1999; 106: 3649–3658. doi: 10.1121/1.428217 PMID: 10615703
14. Iverson P, Hazan V, Bannister K. Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. J Acoust Soc Am. 2005; 118: 3267–3278. doi: 10.1121/1.2062307 PMID: 16334698
15. Lively SE, Pisoni DB, Yamada RA, Tohkura Y, Yamada T. Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. J Acoust Soc Am. 1994; 96: 2076–2087. doi: 10.1121/1.410149 PMID: 7963022

16. Bradlow AR, Akahane-Yamada R, Pisoni DB, Tohkura Y. Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. J Acoust Soc Am. 1999; 61: 977–985.

17. Hirata Y, Whitehurst E, Cullings E. Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. J Acoust Soc Am. 2007; 121: 3837–3845. doi: 10.1121/1.2734401 PMID: 17552731

18. Lively SE, Logan JS, Pisoni DB. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. J Acoust Soc Am. 1993; 94: 1242–1255. doi: 10.1121/1.408177 PMID: 8408964

19. Bradlow AR, Pisoni DB, Akahane-Yamada R, Tohkura Y. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. J Acoust Soc Am. 1997; 101: 2299–2310. doi: 10.1121/1.418276 PMID: 9104031

20. Petersen AM, Tenenbaum JN, Havlin S, Stanley HE, Perc M. Languages cool as they expand: Allometric scaling and the decreasing need for new words. Sci Rep. 2012; 2. 943. doi: 10.1038/srep00943 PMID: 23230508

21. Fought JG, Munroe RL, Fought CR, Good EM. Sonority and climate in a world sample of languages. Cross-Cultural Res. 2004; 38: 27–51. doi: 10.1177/1069397103259439

22. Ember CR, Ember M. Climate, econiche, and sexuality: Influences on sonority in language. Am Anthropol. 2007; 109: 180–185. doi: 10.1525/aa.2007.109.1.180

23. Munroe RL, Fought JG, Macaulay RKS. Warm climates and sonority classes: Not simply more vowels and fewer consonants. Cross-Cultural Res. 2009; 43: 123–133. doi: 10.1177/1069397109331485

24. Atkinson QD. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science. 2011; 332: 346–349. doi: 10.1126/science.1199295 PMID: 21493858

25. Hay J, Bauer L. Phoneme inventory size and population size. Language. 2007; 83: 388–400. doi: 10.1353/lan.2007.0071

26. Wichmann S, Rama T, Holman EW. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. Linguist Typology. 2011; 15: 177–197.

27. Jackendoff R. Possible stages in the evolution of the language capacity. Trends Cogn Sci. 1999; 3: 272–279. doi: 10.1016/S1364-6613(99)01333-9 PMID: 10377542

28. Nichols J. Linguistic complexity: a comprehensive definition and survey. In: Sampson G, Gil D, Trudgill P, editors. Language Complexity as an Evolving Variable. Oxford: Oxford University Press; 2009. pp. 110–125.

29. Sinnemäki K. Complexity in core argument marking and population size. In: Sampson G, Gil D, Trudgill P, editors. Language Complexity as an Evolving Variable. Oxford: Oxford University Press; 2009. pp. 126–140.

30. Haspelmath M, Dryer M, Gil D, Comrie B. The World Atlas of Language Structures Online. n.d. Munich: Max Plank Digital Library; Available: http://wals.info/

31. Reali F, Chater N, Christiansen MH. The paradox of linguistic complexity and community size. In: Cartmill, EA, Roberts, S, Lyn, H, Cornish, H, editors. Proceedings of the 10th International Conference on the Evolution of Language (EVOLANG X). Singapore: World Scientific Publishing Co. Pte. Ltd.; 2014. pp. 270–279.

32. Perkins RD. Deixis, Grammar and Culture. Amsterdam: Benjamins; 1992.

33. Martowicz A. The origin and functioning of circumstantial clause linkers: a cross-linguistic study. Ph.D. Thesis, University of Edinburgh. 2011.

34. Maas U. Orality versus literacy as a dimension of complexity. In: Sampson G, Gil D, Trudgill P, editors. Language Complexity as an Evolving Variable. Oxford: Oxford University Press; 2009. pp. 164–177.

35. McWhorter JH. Defining creole. Oxford: Oxford University Press; 2005.

36. Trudgill P. Linguistic and social typology: The Austronesian migrations and phoneme inventories. Linguist Typology. 2004; 8: 305–320. doi: 10.1515/lity.2004.8.3.305

37. Saffran JR, Newport EL, Aslin RN. Word segmentation: the role of distributional cues. J Mem Lang. 1996; 621: 606–621. doi: 10.1006/jmla.1996.0032

38. Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996; 274: 1926–1928. doi: 10.1126/science.274.5294.1926 PMID: 8943209

39. Johnson EK, Jusczyk PW. Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. J Mem Lang. 2001; 44: 548–567. doi: 10.1006/jmla.2000.2755

40. Frank MC, Tenenbaum JB, Gibson E. Learning and long-term retention of large-scale artificial languages. PLoS One. 2013; 8: e52500. doi: 10.1371/journal.pone.0052500 PMID: 23300975

41. Weiss DJ, Gerfen C, Mitchel AD. Speech segmentation in a simulated bilingual environment: a challenge for statistical learning? Lang Learn Dev. 2014; 5: 30–49. doi: 10.1080/15475440802340101

42. Saffran JR, Johnson EK, Aslin RN, Newport EL. Statistical learning of tone sequences by human infants and adults. Cognition. 1999; 70: 27–52. doi: 10.1016/S0010-0277(98)00075-4 PMID: 10193055

43. Fiser J, Aslin RN. Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. Psychol Sci. 2001; 12: 499–504. doi: 10.1111/1467-9280.00392 PMID: 11760138

44. Kirkham NZ, Slemmer JA, Johnson SP. Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition. 2002; 83: B35–B42. doi: 10.1016/S0010-0277(02)00004-5 PMID: 11869728

45. Hauser MD, Newport EL, Aslin RN. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. Cognition. 2001; 78: B53–B64. doi: 10.1016/S0010-0277(00)00132-3 PMID: 11124355

46. Frank MC, Goldwater S, Griffiths TL, Tenenbaum JB. Modeling human performance in statistical word segmentation. Cognition.; 2010; 117: 107–125. doi: 10.1016/j.cognition.2010.07.005 PMID: 20832060

47. Dutoit T, Pagel V, Pierret N, Bataille F, van der Vrecken O. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. Proceedings of the Fourth International Conference on Spoken Language. 1996. pp. 1393–1396.

48. Thiessen ED, Saffran JR. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Dev Psychol. 2003; 39: 706–716. doi: 10.1037/0012-1649.39.4.706 PMID: 12859124

49. Yang J, Li P. Brain networks of explicit and implicit learning. PLoS ONE. 2012; 7: e42993. doi: 10.1371/journal.pone.0042993 PMID: 22952624

50. Witt A, Puspitawati I, Vinter A. How explicit and implicit test instructions in an implicit learning task affect performance. PLoS ONE. 2013; 8: e53296. doi: 10.1371/journal.pone.0053296 PMID: 23326409

51. Finn AS, Lee T, Kraus S, Hudson Kam CL. When it hurts (and helps) to try: the role of effort in language learning. PLoS ONE. 2014; 9: e101806. doi: 10.1371/journal.pone.0101806 PMID: 25047901

52. Kachergis G, Yu C, Shiffrin RM. Cross-situational word learning is both implicit and strategic. Front Psychol. 2014; 5: 588. doi: 10.3389/fpsyg.2014.00588 PMID: 24982644

53. Arciuli J, Torkildsen J von K, Stevens DJ, Simpson IC. Statistical learning under incidental versus intentional conditions. Front Psychol. 2014; 5: 747. doi: 10.3389/fpsyg.2014.00747 PMID: 25071692

54. Core Team R. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available: http://www.r-project.org/

55. Bates D, Maechler M, Bolker B. lme4: Linear mixed-effects models using S4 classes. 2013. Available: http://cran.r-project.org/package = lme4

56. Barr DJ, Levy R, Scheepers C, Tily HJ. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J Mem Lang; 2013; 68: 255–278. doi: 10.1016/j.jml.2012.11.001

57. Finn A, Hudson Kam CL. Why segmentation matters: Experience-driven segmentation errors impair "morpheme" learning. J Exp Psychol Learn Mem Cogn; 2015 Mar 2. [Epub ahead of print] doi: 10.1037/xlm0000114 PMID: 25730305

58. Kenesei I, Vago RM, Fenyvesi A. Hungarian. London: Routledge; 1998.

59. Siptár P, Törkenczy M. The Phonology of Hungarian. Oxford: Oxford University Press; 2007.

60. Hayes B, Londe ZC. Stochastic phonological knowledge: the case of Hungarian vowel harmony. Phonology. 2006; 23: 59–104. doi: 10.1017/S0952675706000765

61. Ringen CO, Vago RM, Ringen C. Hungarian vowel harmony in optimality theory. Phonology. 1998; 15: 393–416. doi: 10.1017/S0952675799003632

62. Be uš Š, Gafos A, Goldstein L. Phonetics and Phonology of Transparent Vowels in Hungarian. In: Nowak PM, Yoquelet C, Mortensen D, editors. Proceedings of the 3rd Speech Prosody Conference. 2004. pp. 486–497.

63. Connolly JH. Quantifying target-realization differences. Part I: Segments. Clin Linguist Phon. 1997; 11: 267–287. doi: 10.3109/02699209708985196

64. Connolly JH. Quantifying target-realization differences. Part 11: Sequences. Clin Linguist Phon. 1997; 11: 289–298. doi: 10.3109/02699209708985196

65. Sullivan J, McMahon A. Phonetic comparison, varieties, and networks: Swadesh's influence lives on here too. Diachronica. 2010; 27:325–340. doi: 10.1075/dia.27.2.08sul

66. Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. J Mem Lang; 59: 390–412. doi: 10.1016/j.jml.2007.12.005

67. Hudson Kam CL, Newport EL. Getting it right by getting it wrong: when learners change languages. Cogn Psychol; 2009; 59: 30–66. doi: 10.1016/j.cogpsych.2009.01.001

68. Smith K, Wonnacott E. Eliminating unpredictable variation through iterated learning. Cognition; 2010; 116: 444–449. doi: 10.1016/j.cognition.2010.06.004 PMID: 20615499

69. Feher O, Kirby S, Smith K. Social influences on the regularization of unpredictable variation. Proceedings of the 36th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society; 2014. pp. 2187–2191.

70. Barton ME, Tamasello M. The rest of the family: the role of fathers and siblings in early language development. In: Galloway C, Richards BJ, editors. Input and Interaction in Language Acquisition. Cambridge: Cambridge University Press; 1994. pp. 109–134.

71. Boyd R, Richerson PJ. Culture and the Evolutionary Process. Chicago: University of Chicago Press; 1985.

72. Nettle D. Is the rate of linguistic change constant? Lingua. 1999; 108: 119–136. doi: 10.1016/S0024-3841(98)00047-3

73. Dahl Ö. The Growth and Maintenance of Linguistic Complexity. Amsterdam: John Benjamins; 2004.

74. Bentz C, Winter B. Languages with more second language learners tend to lose nominal case. Lang Dyn Chang. 2013; 3: 1–27.

75. Trueswell JC, Sekerina I, Hill NM, Logrip ML. The kindergarten-path effect: studying on-line sentence processing in young children. Cognition. 1999; 73: 89–134. doi: 10.1016/S0010-0277(99)00032-3 PMID: 10580160

76. Snedeker J, Trueswell JC. The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing. Cogn Psychol. 2004; 49:238–299. doi: 10.1016/j.cogpsych.2004.03.001 PMID: 15342261

77. Weighall AR. The kindergarten path effect revisited: children's use of context in processing structural ambiguities. J Exp Child Psychol. 2008; 99: 75–95. doi: 10.1016/j.jecp.2007.10.004 PMID: 18070628

2.3 Experiment 3: input variability and string segmentation

This experiment is an earlier version of Experiment 1 with the same condition-dependent differences in training, as illustrated in Figure 1. In this experiment, however, the participants in both conditions were presented with less training. In the first experiment of Saffran et al. (1996b), the participants were exposed to 3 blocks of approximately 7 minutes of training each. Having shorter training blocks of approximately 4 minutes each was ultimately an attempt to reduce the cost of the participant fees.

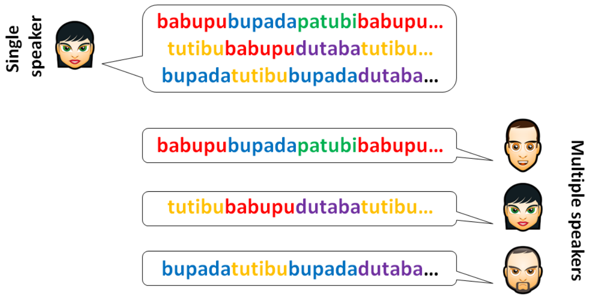

Single speaker
babupubupadapatubibabupu...
tutibubabupudutabatutibu...
bupadatutibubupadadutaba...
babupubupadapatubibabupu...
tutibubabupudutabatutibu...
bupadatutibubupadadutaba...
Multiple speakers

Figure 1: Training regime by condition for Experiments 1 (in Atkinson et al., 2015) and 3.

2.3.1 Materials and methods

The methodology differed from that of Experiment 1 in the following ways:

- The input string was only generated once, so that all participants were presented with the same list of syllables in the same order;
- Only the first 12 minutes of this input string was used, segmented into 3 testing blocks of approximately 4 minutes each;
- The rest period after the first and second testing blocks was 2 minutes (reduced from the 5 minutes in Saffran et al. (1996b));
- The foils were only generated once, so that all participants were presented with the same word-foil testing pairings (though the order in which they were presented was random for each participant).[15]

In Saffran et al. (1996b), the mean score in testing was 76%. As this was substantially above chance performance of 50%, I anticipated that even with reduced training, there would still be evidence of the participants being able to segment the input, and any condition-dependent differences would still be apparent. Subsequent work had also demonstrated that lower levels of training, while correlating with reduced accuracy in testing, could result in successful string segmentation in very similar tasks (Frank et al., 2010).[16] An informal pilot study also suggested that the reduced amount of training was still appropriate.[17]

[15] The randomly-generated foils used in this experiment were *butidu*, *dabiti*, *dibipa*, *pitupa*, *piduba*, and *pitabi*.

[16] Saffran et al. (1996a) themselves had also demonstrated that 8-month-old infants could segment (simpler) speech streams with only 2 minutes of exposure.

[17] The pilot was “informal” in that it involved 7 other Linguistics & English Language Ph.D. and Masters

2.3.2 Participants

48 native English speakers (15 male; aged between 18 and 40, mean 23.6) were recruited at the University of Edinburgh. Each was compensated £3.50. The experiment was run between 23rd June and 17th July 2014.

2.3.3 Analysis and results

The results are shown in Figure 2.

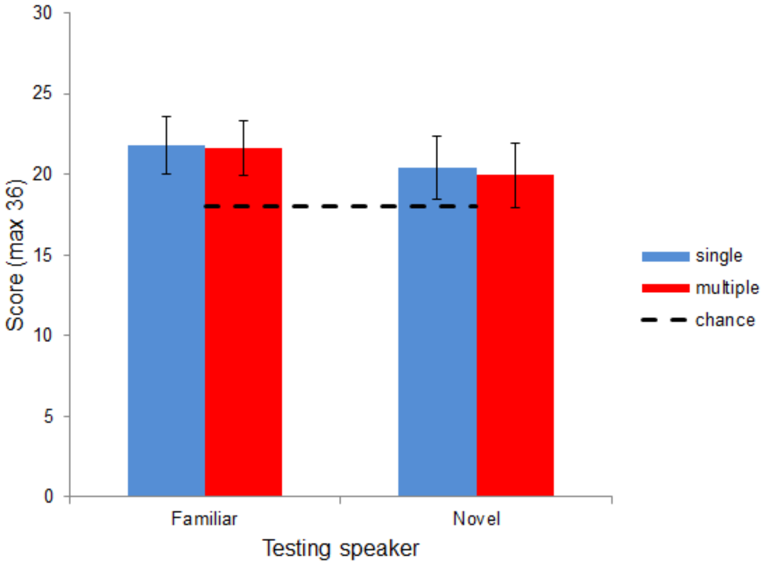


Figure 2: Average scores for each condition in both familiar speaker and novel speaker testing blocks. Error bars are 95% confidence intervals.

A linear mixed effects analysis with score (1 for correct; 0 otherwise) as dependent variable was performed as in Experiment 1. A maximal model with logit regression included *condition* (single speaker or multiple speaker), *speaker identity* (familiar or novel), *order of tests* (familiar speaker test first or second) and the interaction of *condition* and *speaker identity* as (centred) fixed effects, with *participant identity* as a random effect. As in Experiment 1, the interaction of *condition* and *speaker identity* was included to see if there was any effect of participants in the multiple speaker condition being better at distinguishing words from foils when listening to unfamiliar speakers, following similar findings in HVPT (Lively et al., 1993, 1994; Bradlow et al., 1999).

The model (AIC = 4655, BIC = 4692) offered a significantly better fit to the data compared to the equivalent null model (AIC = 4663, BIC = 4675) ($\chi^2(4) = 15.988$, p = 0.003). Under AIC, the model appears to describe the data better than null, although under BIC there is some suggestion that the model is overparamaterised. Using either the model or the null model, the intercept is non-zero ($\beta \geq 0.340$, SE = 0.059 , p <0.001), suggesting that the testing scores

students, who were not all native speakers of English, and who were not paid. These participants would likely have been more familiar with general experimental design and procedure than the average participant, and would have had different motivations for taking part and scoring as highly as possible. 3 were placed in the single-speaker condition, and had an average testing score of 81% (s.d. = 12%). This was greater than that of Saffran et al. (1996b), despite the reduced training times. 4 participants were placed in the multiple-speaker condition, and had an average score of 66% (s.d. = 17%).

are above chance. Under the (non-null) model, there were significant contributions of *speaker identity* (β = -0.205, SE = 0.071, p = 0.004) and *order of tests* (β = -0.222, SE = 0.071, p = 0.002). There were no effects of *condition* (β = 0.038, SE = 0.120, p = 0.752), or the interaction of *condition* and *speaker identity* (β = 0.078, SE = 0.141, p = 0.583).

As in Experiment 1, this analysis suggests that the participants were able to use the distributional cues in their training strings to discriminate between words and non-words, replicating the result of Saffran et al. (1996b). Performance was better in the familiar voice testing: participants were 1.56 times as likely to produce a correct response as incorrect, corresponding to an accuracy of 61%; in the novel voice testing, they were 1.27 times as likely, corresponding to an accuracy of 56%. Greater performance in the familiar voice testing supports the HVPT findings that distinguishing words is easier when they are presented by familiar speakers (Lively et al., 1993, 1994; Bradlow et al., 1999). Scores in the second test blocks were also on average lower than those in the first (on average 56% accurate compared to 60%), suggesting either an effect of participant fatigue or interference from the first block. There is also evidence of the participants being able to generalise their training input to a novel speaker. Considering only the novel voice testing data presented in the first block, a linear mixed model with logit regression and no fixed effects and *participant identity* as a random effect had an intercept significantly greater than zero (β = 0.223, SE = 0.085, p = 0.009): participants were 1.25 times as likely to produce a correct response as incorrect, corresponding to an accuracy of 56%.

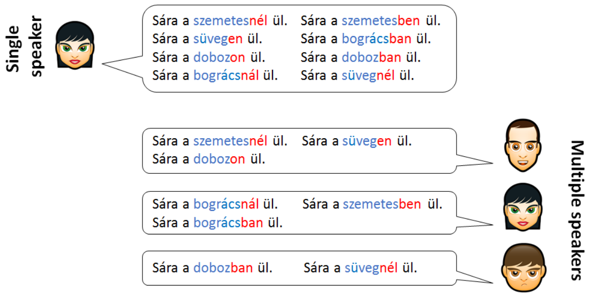
2.3.4 Discussion and conclusion

This experiment replicates the main results of the first experiment of Saffran et al. (1996b) in demonstrating that participants can use distributional cues to segment a continuous speech stream. We also see this is possible when input variability is increased and the input is provided by three voices instead of one, as well as demonstrating that learners can generalise their training input to novel voices.

The performance in testing, however, was somewhat lower than in the original study. Average testing accuracy was 58%, compared to Saffran et al.'s (1996b) 76%. This at least partly appears to be an effect of the extended testing procedure, with half of the testing being provided by a novel speaker and half of the testing taking place in a second testing block. One concern resulting from the lower accuracy scores was that they could have masked any between-condition differences. The experiment design was therefore changed for Experiment 1 in Atkinson et al. (2015), with the amount of training restored to that of Saffran et al. (1996b) (3 blocks of approximately 7 minutes of input, with 5 minute breaks after the first and second blocks). In contrast to Saffran et al. (1996b), and to reduce any influence of the order of a particular input string, 24 different input strings were generated, each independently randomised, and each used only once in each experimental condition. Unique sets of 6 testing foils were also created to be used in testing, so that there was one for each input string. Despite these changes, Experiment 1 essentially replicated the results of this study in finding no evidence for an effect of input variability on the ability to segment continuous speech streams.

2.4 Experiment 4: input variability and morphological acquisition

This experiment is an earlier version of Experiment 2 with the same condition-dependent differences in training, as illustrated in Figure 3. This version had a smaller sample size of 10 participants in each condition, rather than 20. The lack of difference between the conditions evident in Experiment 2 was not quite so clear here, however, hence the replication with a larger sample size.



Sára a szemetesnél ül. Sára a szemetesben ül.
Sára a süvegen ül. Sára a bográcsban ül.
Sára a dobozon ül. Sára a dobozban ül.
Sára a bográcsnál ül. Sára a süvegnél ül.

Sára a szemetesnél ül. Sára a süvegen ül.
Sára a dobozon ül.

Sára a bográcsnál ül. Sára a szemetesben ül.
Sára a bográcsban ül.

Sára a dobozban ül. Sára a süvegnél ül.

Figure 3: Training regime by condition for Experiments 2 (in Atkinson et al. (2015)) and 4.

2.4.1 Materials and methods

The methodology for this experiment only differed from that of Experiment 2 in that it had an additional reception test at the end. This was a forced-choice judgement task of aurally-presented stimuli, designed to be an additional test of participant acquisition of the morphological system. This test was replaced in Experiment 2 by a post-test interview with each participant, which is described in Section 2.5 below.

In the reception test, the participants were tested on their ability to discriminate between sentences of the target language and incorrect sentences. A fourth native speaker of Hungarian (and trained phonetician) recorded 3 sets of labels for the 12 stimuli. One was the target language. The other two were sets of foils: in one set, each sentence had an incorrect alternation as the noun suffix; in the other, a vowel was changed in the noun stem. The different types of foils were intended to discriminate between errors specifically related to the morphology of the language and those relating to more general acquisition difficulties. The target language is replicated in Table 1 for ease of comparison, and the sets of foils given in Tables 2 and 3.

Table 1: Target language for Experiments 2 and 4.

	Foil	Noun vowels
1.	Sára a süvegben ül.	/y/ /ɛ/ /ɛ/
2.	Sára a süvegnél ül.	/y/ /ɛ/ /eː/
3.	Sára a süvegen ül.	/y/ /ɛ/ /ɛ/
4.	Sára a szemetesben ül.	/ɛ/ /ɛ/ /ɛ/ /ɛ/
5.	Sára a szemetesnél ül.	/ɛ/ /ɛ/ /ɛ/ /eː/
6.	Sára a szemetesen ül.	/ɛ/ /ɛ/ /ɛ/ /ɛ/
7.	Sára a dobozban ül.	/o/ /o/ /ɔ/
8.	Sára a doboznál ül.	/o/ /o/ /aː/
9.	Sára a dobozon ül.	/o/ /o/ /o/
10.	Sára a bográcsban ül.	/o/ /aː/ /ɔ/
11.	* Sára a bográcsnál ül.	/o/ /aː/ /aː/
12.	Sára a bográcson ül.	/o/ /aː/ /o/

Table 2: Foils with grammatical errors for Experiment 4 reception test. Each suffix is the incorrect vowel-harmony-dependent alternation, marked in **bold**.

	Foil	Noun vowels
1.	Sára a süveg**ban** ül.	/y/ /ɛ/ /ɔ/
2.	Sára a süveg**nál** ül.	/y/ /ɛ/ /a:/
3.	Sára a süveg**on** ül.	/y/ /ɛ/ /o/
4.	Sára a szemetes**ban** ül.	/ɛ/ /ɛ/ /ɛ/ /ɔ/
5.	Sára a szemetes**nál** ül.	/ɛ/ /ɛ/ /ɛ/ /a:/
6.	Sára a szemetes**on** ül.	/ɛ/ /ɛ/ /ɛ/ /o/
7.	Sára a doboz**ben** ül.	/o/ /o/ /ɛ/
8.	Sára a doboz**nél** ül.	/o/ /o/ /e:/
9.	Sára a doboz**en** ül.	/o/ /o/ /ɛ/
10.	Sára a bogrács**ben** ül.	/o/ /a:/ /ɛ/
11.	Sára a bogrács**nél** ül.	/o/ /a:/ /e:/
12.	Sára a bogrács**en** ül.	/o/ /a:/ /ɛ/

Table 3: Foils with incorrect vowels in the noun stems for Experiment 4 reception test. Vowels changed (to other Hungarian vowels) are shown in **bold**.

	Foil	Noun vowels
1.	Sára a süv**a**gben ül.	/y/ /ɔ/ /ɛ/
2.	Sára a s**i**vegnél ül.	/i/ /ɛ/ /e:/
3.	Sára a süv**ö**gen ül.	/y/ /ø:/ /ɛ/
4.	Sára a sz**i**metesben ül.	/i/ /ɛ/ /ɛ/ /ɛ/
5.	Sára a szem**a**tesnél ül.	/ɛ/ /ɔ/ /ɛ/ /e:/
6.	Sára a sz**é**metesen ül.	/e:/ /ɛ/ /ɛ/ /ɛ/
7.	Sára a d**ö**bozban ül.	/o:/ /o/ /ɔ/
8.	Sára a dob**ö**znál ül.	/o/ /o:/ /a:/
9.	Sára a d**u**bozon ül.	/u/ /o/ /o/
10.	Sára a bogr**e**csban ül.	/o/ /ɛ/ /ɔ/
11.	Sára a b**a**grácsnál ül.	/ɔ/ /a:/ /a:/
12.	Sára a bogr**ú**cson ül.	/o/ /u:/ /o/

The participants were presented with all 36 utterances in a random order alongside the corresponding visual stimuli. After each, they were required to judge whether they had just heard a "correct" sentence in the language they had been trying to learn, or whether there was an error, be it of pronunciation or grammar.

2.4.2 Participants

20 native English speakers (8 male; aged between 19 and 49, mean 24.3) were recruited at the University of Edinburgh. Each was compensated £10. The experiment was run between 5th and 29th August 2013. Data for one original participant was excluded as they failed to turn up for the third day, and a new participant was recruited in their place.

2.4.3 Analysis and results

The analysis was carried out as for Experiment 2.

2.4.3.1 Production of the noun stems

Mean stem accuracy for each of the conditions over the 6 rounds is illustrated in Figure 4.

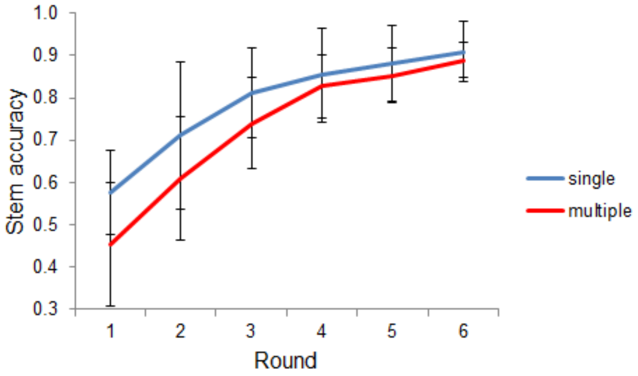


Figure 4: Accuracy of participant productions of target stems. Error bars are 95% confidence intervals.

A linear mixed effects analysis included *condition* (single speaker or multiple speaker), *novelty* (whether the target stimulus had been seen in training or not) and *round* and their interactions as (centred) fixed effects. *Participant identity* was investigated as a random effect. This model was significantly better than the equivalent null model ($\chi^2(7) = 384.31$, p <0.001). There were significant effects of *round* ($\beta = 0.076$, SE = 0.004, t(1433) = 20.437, p <0.001), and *novelty* ($\beta = 0.033$, SE = 0.013, t(1433) = 2.454, p = 0.014), but no effect of *condition* ($\beta =$ 0.068, SE = 0.053, t(1433) = 1.286, p = 0.199). There was a significant effect of the interaction of *condition* and *round* ($\beta =$ -0.025, SE = 0.007, t(1433) = -3.366, p $\leq$ 0.001), with accuracy in the multiple speaker condition increasing to a greater extent than in the single speaker. There was no effect of any of the other interaction terms ($|\beta| \leq 0.025$, SE $\geq$ 0.008, $|$t(1433)$| \leq 1.226$, p $\geq$ 0.220).

2.4.3.2 Production of the affixes

The results by condition for each measure are shown in Figure 5. Average scores for the whole language are given, along with a comparison of the production accuracy for the stimuli seen in training and novel items.

Linear mixed analyses for each measure used logit regression and models which again included *condition*, *novelty* and *round* and their interactions as (centred) fixed effects. *Participant identity* was again included as a random effect. For all three measures, the fitted model was better than the corresponding null model (case identification: $\chi^2(7) = 183.97$, p <0.001; case accuracy: $\chi^2(7) = 120.54$, p <0.001; alternation accuracy: $\chi^2(7) = 100.93$, p <0.001). The results of the models are summarised in Table 4.

2.4.3.3 Reception test

Average success score, measured by correct acceptance of correct sentences and correct rejection of incorrect sentences was 52% for participants in the single-speaker condition and 60% for participants in the multiple-speaker. To control for differences in individual participant tendencies to accept or reject a test item, I employed a *d'* analysis to compare responses between

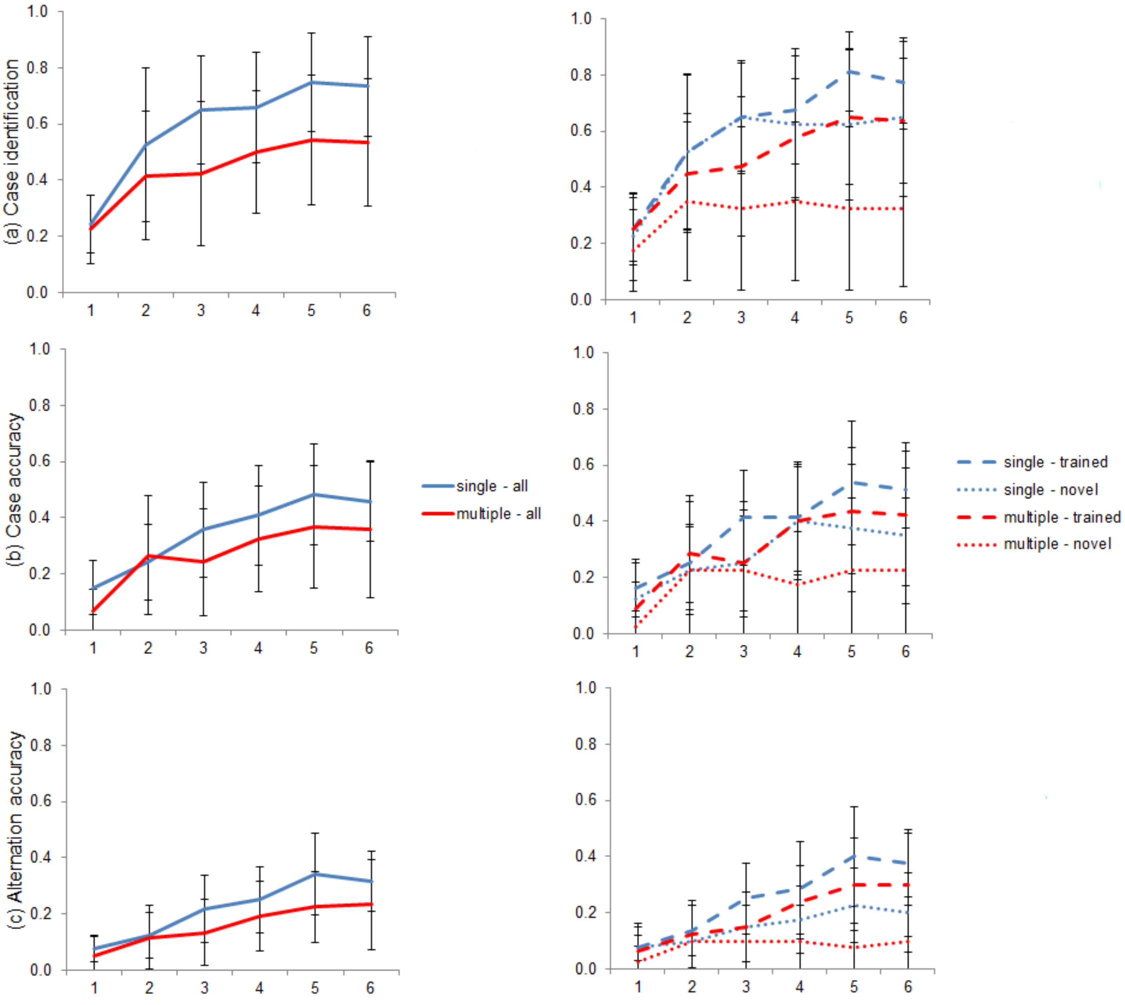


Figure 5: Acquisition of the suffixes. Mean scores by condition are shown for the 3 measures, both for the entire target set (left), and split by training and novel image labels. Error bars are 95% confidence intervals.

conditions.[18] Hit rate was 0.62 ± 0.20 (95% confidence interval) in the single-speaker condition and 0.65 ± 0.14 in the multiple speaker. False alarm rate was 0.53 ± 0.13 in the single-speaker and 0.43 ± 0.09 in the multiple. There was no difference in the *d'* scores between the conditions (F(1,18) = 2.372, p = 0.141).

2.4.4 Discussion and conclusion

As in Experiment 2, stem acquisition improved with increased training and testing, and produced stems were more accurate for images seen in training than for novel images. In this

[18] *d'* is calculated by

$$d' = Z(\textit{Hit rate}) - Z(\textit{False alarm rate})$$

where Z(p) is the inverse cumulative Gaussian distribution for $p \in (0, 1)$, *Hit rate* is the proportion of correct test items accepted, and *False alarm rate* is the proportion of foils accepted. *d'* estimates are not defined for *Hit rate* or *False alarm rate* scores of 1 or 0, but no 1 or 0 scores were encountered in this sample and so no adjustments needed to be made.

Measure	Fixed effect	β	SE	Pr(>|z|)	
Case identification	condition	1.024	0.690	0.138	
	novelty	0.807	0.140	<0.001	***
	round	0.433	0.041	<0.001	***
	condition*novelty	-0.803	0.281	0.004	**
	condition*round	0.170	0.082	0.037	*
	novelty*round	0.270	0.083	0.001	**
	condition*novelty*round	-0.108	0.165	0.515	
Case accuracy	condition	0.854	0.654	0.191	
	novelty	0.723	0.150	<0.001	***
	round	0.366	0.041	<0.001	***
	condition*novelty	-0.457	0.301	0.129	
	condition*round	-0.023	0.083	0.782	
	novelty*round	0.152	0.088	0.085	.
	condition*novelty*round	-0.104	0.176	0.555	
Alternation accuracy	condition	0.683	0.498	0.170	
	novelty	0.822	0.179	<0.001	***
	round	0.331	0.047	<0.001	***
	condition*novelty	-0.376	0.358	0.294	
	condition*round	0.048	0.093	0.605	
	novelty*round	0.221	0.105	0.035	*
	condition*novelty*round	-0.114	0.209	0.587	

Table 4: Summary of linear mixed modelling results for the three affix production measures.

experiment, the participants in the multiple-speaker condition improved more than in the single-speaker condition, eliminating the difference in production accuracy in the earlier rounds that can be seen in Figure 4. There is no effect of condition at the model intercept, however, again suggesting that input variability does not affect language acquisition in general.

In the suffixes, we see evidence of increased training and testing improving acquisition, and of trained items being easier to produce than novel items. This mirrors the findings of Experiment 2, but from two of our measures (case identification and alternation accuracy) we also see some evidence of a greater improvement in the production of trained items over novel.

In Experiment 2, there was no suggestion that condition had any effect on the acquisition of the stems on any of the 3 measures, either as an isolated variable or as part of an interaction. Here, however, there is some indication of a condition-dependent difference in performance. On the case identification measure, participants in the single-speaker condition outperform those in the multiple-speaker overall, with the majority of that effect coming from the labelling of novel items. This would lend support to the hypothesis that increased input variability does hamper the acquisition of complex morphology (Nettle, 2012). Learning from a greater number of speakers appears to have made generalisation from the training data particularly difficult.

Considering the more stringent measures of case accuracy and alternation accuracy, however, there is no evidence of an effect of condition. Nor is there any evidence of single-speaker conditioned participants being able to better distinguish between target language sentences and foils in the reception task. With these observations, and noting the smaller sample size of this experiment, a replication with a larger sample size was deemed necessary.

A further question came from looking at the incorrect productions for the novel stimuli.

Many of the novel images were labelled with one of the training labels for an image involving the same container, often with a very high degree of accuracy. Sentences 1, 2 and 3 in the target language (see Table 1), for example, all describe the position of the mouse relative to the hat. Where a participant was trained on Sentences 1 and 2 and so encountered the image for Sentence 3 as a novel item in testing, they often labelled it with either Sentence 1 or Sentence 2, rather than generalising from their training data and producing Sentence 3.

We can quantify the extent to which participants used a novel label for a novel image. For each participant, I considered each of the labels for a given container (hat, basket, box and cauldron) in turn, giving a score of 1 if the novel testing image was marked with a suffix which did not match either of the ones used for either of the images seen in training, and 0 otherwise. I then averaged the score across the 4 containers for each participant. A score of 0.75, for example, would then indicate that 3 of the novel images were labelled with unique suffixes, and 1 was labelled with one of the suffixes also used for (at least) one of the other images involving the same container. Across all participants and rounds, the average unique labelling of novel images was only 0.60 ± 0.06 (95% confidence intervals). Even at Round 6, it was 0.85 ± 0.17 for participants in the single-speaker condition, but a much lower 0.43 ± 0.28 in the multiple-speaker condition.

It was not immediately clear why participants would reuse training labels and apply them to novel stimuli. They could have (a) failed to segment the training data, (b) successfully segmented the training data but failed to link the segments to meanings, or (c) succeeded in segmenting the data and linking the segments to meanings, but not used this knowledge to construct novel sentences for some other reason. It was also not clear whether or not the participants in such cases knew that they were applying a label they were trained on to describe a novel images. Experiment 2 therefore concluded with a brief interview with the participant. This additional material was mentioned briefly in the PLoS paper, but I describe it in more detail in the following section.

2.5 Supplementary material for Experiment 2

In addition to the details described in Atkinson et al. (2015), Experiment 2 included a post-test interview with each participant, both as an additional assessments of the acquisition of the morphological systems, and to try and find out why participants appeared to be so regularly reapplying training images to novel images, rather than producing more appropriate novel labels.

The participants were interviewed immediately after the Round 6 production test of Experiment 2.[19] They were asked the following questions:

1. Which language do you think this is?
2. You were asked to describe pictures which you weren't trained on the description for.
   - Did you realise this?
   - What did you do? (Why?)
3. How do you say "hat" in the language?

[19] These interviews took place in the recording booth of the main experiment and were recorded. Verbal permission to record the interviews was obtained from the participants before the interview started.

4. How do you say "basket" in the language?[20]
5. How do you say "box" in the language?
6. How do you say "cauldron" in the language?
7. How do you say "in" in the language?
8. How do you say "next to" in the language?
9. How do you say "on top of" in the language?
10. Is there only one way to say "in", "next to" and "on top of"?

The questions were repeated and reworded as necessary to ensure the participants understood what was being asked of them. Further clarification questions were also asked when necessary.

The multiple-speaker participants were able to provide unambiguous labels for 2.00 ± 0.82 (95% confidence interval) of the 4 stems on average, compared to 2.50 ± 0.75 in the single-speaker condition. The difference is not significant (Welch's two-sample t-test, t(37.733) = 0.942, p = 0.352). For the 3 suffixes, the multiple-speakers were able to produce unambiguous labels for 1.60 ± 0.62, compared to 1.90 ± 0.57 in the single-speaker condition. Again, this is not significant (t(37.729) = 0.875, p = 0.387).

16 (80%) of the participants in the multiple-speaker condition reported that they realised that there were novel stimulus items to describe in the testing stages, compared to only 10 (50%) of the single-speaker participants. Of these, 12 of the multiple-speaker participants (75% of 16) said they attempted to produce novel utterances to describe the novel stimuli, as opposed to reproducing one of the training utterances, compared to 6 (60% of 10) in the single-speaker condition. Of those that did not attempt to produce a novel label, each put this down to having been unable to segment the training input well enough. Reusing a training label was understood to be inaccurate, but felt to be the most appropriate thing they could do.

All (100%) of the multiple-speaker participants and all but one (95%) of the single-speaker said that they had attempted to parse the training utterances to determine the underlying grammatical system of the target language during the experiment. 8 (40%) multiple-speaker compared to 10 (50%) single-speaker participants appeared to have convincingly and consistently succeeded in segmenting the noun stems and the affixes.

10 (50%) of the participants in the multiple-speaker condition appeared to have been aware of some semantically-redundant variability in the case marking, compared to 12 (60%) in the single-speaker. Of these, 8 (80%) of the multiple-speaker participants believed that these alternations were systematically conditioned on the nouns, as opposed to resulting from natural speaker variation (either within or between speaker), compared to 10 (83%) in the single-speaker condition. Although some participants had successfully acquired the alternation for one or, more rarely, two of the cases, there were no participants who showed any sign of having fully acquired the vowel-harmony dependent elements of the morphological system.

Overall, the interview data supports the conclusions of the main experiment. The lack of a difference between the conditions when asked to produce the stems and suffixes out of context supports there being no evidence of input-variability effects on the acquisition of the

[20] Although *szemetes* is better translated as "bin" and the image chosen for this container was a wastepaper bin, many of the participants felt that the image would be better described in English as a "basket". To save unnecessary confusion, the participants were asked to give the term for "basket".

morphological system.[21] Failure to produce accurate labels for novel stimuli also appears to have been due to an inability to segment the training input. There is no evidence to suggest that input variability has an effect on this segmentation.

2.6 Conclusion

As discussed in Atkinson et al. (2015, p. 14-17) in Section 2.2, these experiments together offer no support for the proposal that speaker input variability affects the acquisition of morphology, and so no support for such variability being a casual explanation for the link between number of speakers and morphological complexity (Lupyan and Dale, 2010; Nettle, 2012).

[21] Ability to produce these segments was also highly correlated with Round 6 case identification scores ($r=0.71$, $t = 6.21$, $p < 0.001$).

3 Adult learning

3.1 Introduction

Adult learning has been widely proposed as a factor in the simplification of language (Dahl, 2004; Trudgill, 2004a, 2011; McWhorter, 2005; Wray and Grace, 2007) and may be a mechanism by which number of speakers determines linguistic complexity (Lupyan and Dale, 2010; Dale and Lupyan, 2012; Nettle, 2012). This chapter investigates its effects in three experiments.

Adult language acquisition is substantially different from that of children. Though the existence of a critical period remains controversial, there is a clear relationship between age of acquisition and ultimate language proficiency, and strong evidence to suggest that learning a language after puberty leads to both productive and receptive deficiencies, particularly in relation to phonology, morphology, and syntax (Johnson and Newport, 1989; Newport, 1990, 2002; Scovel, 2000; Clark, 2003; Trudgill, 2011). Adult learners find certain linguistic features particularly challenging to acquire, including morphological complexity, syntagmatic and paradigmatic redundancy, and irregularities, even when similar features are found in their native language (Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011). Native-language features also influence the language learning, with phonological, syntactic, lexical and pragmatic transfer. Developmental errors differ from children's, and ultimate attainment is variable, dependent on age of acquisition, learning context, and learner motivation (Selinker, 1972; Bley-Vroman, 1989; Csizér and Dörnyei, 2005; Nettle, 2012).

Languages with greater degrees of adult contact and learning, therefore, are thought to be under an increased pressure for simplification due to these acquisition difficulties. The languages then adapt to the needs and abilities of these older learners, with the more "difficult" features filtered out (Wray and Grace, 2007; Lupyan and Dale, 2010; Bentz and Winter, 2013). There will be fewer "conventional strategies for encoding semantic distinctions such as situational/epistemic possibility, evidentiality, the optative, indefiniteness, future tense, distance contrasts in demonstratives and remoteness distinctions in the past tense" (Lupyan and Dale, 2010, p. 6). Many of these features may be informationally redundant (the information being retrievable from either linguistic or pragmatic context) and so such languages can comfortably tolerate these lower levels of complexity (Dahl, 2004; Gil, 2009; Lupyan and Dale, 2010). Lower complexity will, however, result in a greater reliance on extralinguistic, pragmatic, information, but again it is claimed that this better suits adult learners compared to children (Lupyan and Dale, 2010).

The inverse relationship between number of speakers and morphological complexity can therefore be explained by a correlation between the number of speakers and the proportion of non-native speakers (Lupyan and Dale, 2010). The proportion of non-native speakers would therefore be a more direct determinant of linguistic complexity. We could predict that languages spoken by smaller numbers of speakers, but which have higher proportions of non-native speakers, will have lower levels of linguistic complexity. Languages spoken by larger groups, but with relatively few non-native speakers, will be more complex. Bentz and Winter (2013) offer some support for this. They demonstrate that case marking is negatively correlated to the proportion of second language speakers of a language. They consider a sample of 66 languages, covering 26 language families and 16 geographic areas, and conclude that languages in which the majority of the speakers are second language speakers are those which have no case marking.

As Thurston (1994, p. 603) notes, "simplification needs an opposing force to restore complexity: otherwise all languages would eventually become simple and we would need to explain how complexity arose in the first place". One proposal is that the absence of contact and adult learning creates the ideal, esoteric, circumstances for "mature", complex, language features to develop (Thurston, 1994; Dahl, 2004; Wray and Grace, 2007; Trudgill, 2011). I discuss the mechanisms behind this process in the next chapter (see Section 4.1.) An alternative explanation is that the increased redundancy in more complex languages aids child learning (Lupyan and Dale, 2010). As infants are less competent than adults at using context and extra-linguistic cues to aid comprehension (Trueswell et al., 1999; Snedeker and Trueswell, 2004; Weighall, 2008), increased redundancy may make up for this shortfall by, for example, providing additional cues as to how to segment utterances (Thiessen et al., 2005; Nettle, 2012). Therefore languages which are more complex due to higher levels of redundancy are more likely to be found in populations where most of the learners are children. Agglutinative languages in particular should suit child learning as they are particularly likely to be over-specified, with features such as remoteness being obligatorily marked on the verb (Lupyan and Dale, 2010).[22] Such hypotheses certainly require further investigation, but should be testable (Nettle, 2012).

Returning to simplification, even if adult learners do acquire and produce a more simple morphological system, there still needs to be some mechanism by which this individual-level simplification affects the population-level characteristics of the language. There is a "problem of linkage" (Kirby, 1999).[23] Even if the social conditions result in the children of non-native learners learning the simpler language of their parents (which may be unlikely, given that children typically acquire the language of their speech community), the simplifications will remain restricted to a subset of the population, rather than having an influence on the language as a whole. Children have been shown to eliminate variability in their input (Hudson Kam and Newport, 2009); if they do this when they receive a mix of native and non-native speaker input, then the cross-generational transmission of the complex or simple forms may depend on their frequency in the input, or sociolinguistic factors (such as the relative importance for language acquisition of each of the different individuals who provide the input, Barton and Tomasello, 1994).

In addition, or alternatively, there may be an effect of the interaction between native and non-native speakers. Native speakers may accommodate to non-native speakers by simplifying their language, and these simplifications will then be part of the ambient environment from which child learners receive their linguistic input. There is also the possibility that, through sufficient exposure to the simpler language of non-native speakers and experience of using simplified, Foreigner Directed Speech, the native speakers will also incorporate simplifications into their own language. With enough experience, native speakers may use simplified language features amongst themselves, and so these simplifications may spread through horizontal transmission.

In the following experiments, I investigate the claim that adult learning can lead to morphological simplification. In Experiment 5, I consider the adult learning of a complex morphological system by training and testing adult participants on their acquisition of an artificial language. I

[22] As noted on p. 18, Lupyan and Dale (2010) stress the importance of redundancy as a characteristic of more complex languages and how it may assist child learning. This explanation would appear to have little directly to say about other complex features, such as a greater number of irregularities, though they may be a subsequent consequence of the greater use of morphological strategies over lexical (Jackendoff, 1999; Trudgill, 2011).

[23] More explicitly, the problem of linkage is formulated: "Given a set of observed constraints on cross-linguistic variation, and a corresponding pattern of functional preference, an explanation of this fit will solve the problem: how does the latter give rise to the former?" (Kirby, 1999, p. 20).

find support for the claim that adults initially acquire and produce a simpler system than that of their target language. I then use the data produced in this study as input data for a second set of learners in Experiment 6. Using simpler and more complex subsets of the naturalistic data produced by the participants in Experiment 5 in different combinations, I investigate how the simplifications affect the language through learning in four different types of populations. Viewing complex input data as representative of "native" language and the simpler data as that of "adult learners", I consider how the simplifications may have different effects in larger and smaller groups, and in groups with higher and lower proportions of non-native speakers. Despite the differences in the input data across the different conditions, I ultimately find no difference in the complexity of the morphological systems which are acquired. Hence there is no evidence to explain how simplifications resulting from adult learning could lead to simplifications at the level of the group.

Experiment 7 goes beyond language learning to also consider language *use*, to see if interactions between native and non-native speakers can provide any insights into how individual-level simplification could affect population-level language characteristics. A speaker who acquires a more complex language than their partner simplifies their output in interaction, and so I argue that native speaker accommodation to non-natives may be a key linking mechanism for adult learning affecting language complexity.

3.2 Experiment 5: adult learning and morphological simplification

This experiment has two aims. First, it assesses the claim that non-native speakers will acquire and produce a simpler morphological system than that of their target language. Secondly, it creates a set of linguistic stimuli in preparation for Experiment 6, which assesses the effects of different social-group dependent learning contexts on the transmission of morphological complexity.

3.2.1 Materials and methods

Participants were required to learn an artificial language which describes the 18 static images shown in Figure 6. There are three different animals (a crocodile, a duck, and a bird), pictured either singly or in pairs, and marked with arrows to indicate three different motions. The three dimensions of this meaning space are explicitly laid out in Table 5, alongside the target language. Each image is uniquely described by three words, with "quantifiers" to describe Number, "nouns" to describe Animal and "verbs" to describe Motion. Each word is made up of a stem, shown in black, and a suffix, in red. The quantifier stems were labelled either *won* ("one" for 1) or *sum* ("some" for 2). The noun stems were *snap* (crocodile), *kwak* (duck) or *twit* (bird). The verb stems were *woosh* (straight motion), *boing* (bouncing) or *loop* (looping). These stems were designed to be "pseudo-English", with the expectation that this would make them relatively easy for the participants to acquire (Fehér et al., 2014; Culbertson and Newport, 2015).[24] Participants would then be able to produce sets of labels which unambiguously identify all of the images of the meaning space after only limited training, even if the stems were not acquired with complete accuracy, and regardless of how well the suffixes had been learned. The morphological systems expressed by the suffixes can then be assessed for complexity, and

[24] English itself was not used to limit the impact of English morphology. If "duck" was used in place of "kwak", for example, a plural marker of "-s" may have been anticipated, instead of the target "-op".

the claim that adult learners reduce the complexity specifically arising from largely redundant morphological features (Lupyan and Dale, 2010) assessed.

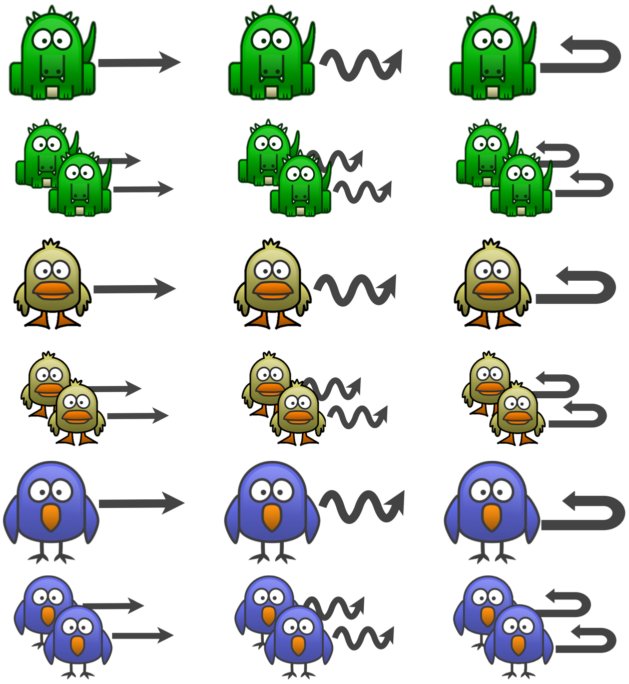


Figure 6: Stimuli set. Made up of every combination of 3 Animals (crocodile, duck and bird), 2 Numbers (1 or 2), and 3 Movements (a straight motion, bouncing and looping).

The suffixes were designed to be dependent on the meaning space to varying degrees, i.e. the set was to be made up of both more regular and more irregular forms. For the quantifiers, -*a* and -*ak* are applied to singular and plural images of ducks and birds; -*u* and -*uk* are the singular and plural suffixes for the crocodile. For the nouns, -*o* and -*op* mark singular and plural, except for ducks, where -*o* is used for both singular and plural. For the verbs, -*an* and -*asp* are used for singular and plural for the duck and bird images, while -*en* and -*esp* are used for the crocodile. An exception arises where the looping motion occurs with a plural image: all animals take the -*onk* suffix.

Participants were required to learn the language over 8 rounds of training and testing. In a training round, 9 of the 18 stimuli and their labels were randomly selected as training items. The participant was then trained on 5 randomly-sorted passes of this set. The image was presented for 1 second, before the image and label text were presented together for a further 5 seconds. The label was then removed, and the participant required to retype the label from memory. Advance to the next training item was controlled by the enter key. No feedback was given as to the accuracy of the retyping.

In a testing round, the participant was presented with all 18 images in a random order and required to provide a label for each. It was not possible to advance to the next image without providing a label.

The experiment was written and run in Matlab (R2010a) with the Psychtoolbox extensions (Brainard, 1997; Kleiner et al., 2007).

	Number	Animal	Motion	Quantifier	Noun	Verb
1.	1	crocodile	straight	wonu	snapo	wooshen
2.	2	crocodile	straight	sumuk	snapop	wooshesp
3.	1	crocodile	bounce	wonu	snapo	boingen
4.	2	crocodile	bounce	sumuk	snapop	boingesp
5.	1	crocodile	loop	wonu	snapo	loopen
6.	2	crocodile	loop	sumuk	snapop	looponk
7.	1	duck	straight	wona	kwako	wooshan
8.	2	duck	straight	sumak	kwakop	wooshasp
9.	1	duck	bounce	wona	kwako	boingan
10.	2	duck	bounce	sumak	kwakop	boingasp
11.	1	duck	loop	wona	kwako	loopan
12.	2	duck	loop	sumak	kwakop	looponk
13.	1	bird	straight	wona	twito	wooshan
14.	2	bird	straight	sumak	twito	wooshasp
15.	1	bird	bounce	wona	twito	boingan
16.	2	bird	bounce	sumak	twito	boingasp
17.	1	bird	loop	wona	twito	loopan
18.	2	bird	loop	sumak	twito	looponk

Table 5: Meaning space and target language. The images are shown in Figure 6. Stems are shown in black; suffixes in red.

A pilot study was run to determine the appropriateness of the experimental design, target language and stimuli, and to determine the appropriate number of rounds of training and testing. Full details can be found in Appendix B.

3.2.2 Participants

26 native English speakers (11 male; aged between 18 and 29, mean 21.6) were recruited at the University of Edinburgh. Each was compensated £7. The experiment was run between 19th May and 13th June 2014.

3.2.3 Analysis and results

3.2.3.1 Acquisition of stems and suffixes

For each round, the stems and suffixes in the productions were isolated. Where the start of each word exactly matched the stem of the target language, this was done automatically. Otherwise this was done by hand. Acquisition of the stems and suffixes was considered separately. For each, the normalised Levenshtein distance was calculated between each production and its target, and an accuracy score defined as 1 minus this distance. The overall accuracy score for each round was then obtained by taking the average accuracy score for each of the 18 stems produced for each of the 3 words of each label. These accuracy scores are illustrated in Figure 7.

Stem (t(25) = -6.861, p <0.001) and suffix accuracy (t(25) = -15.651, p <0.001) were both significantly greater in Round 8 compared to Round 1. Stem accuracy is close to ceiling from Round 4 (average $\geq$ 0.96 from Round 4 onwards). Suffix accuracy is lower, though the productions are close to the target language suffixes by Round 6 (average $\geq$ 0.91 from Round 6 onwards).

I also considered the stability of the suffixes from round to round. This was calculated as the average of 1 minus the normalised Levenshtein distances between each suffix and the previous

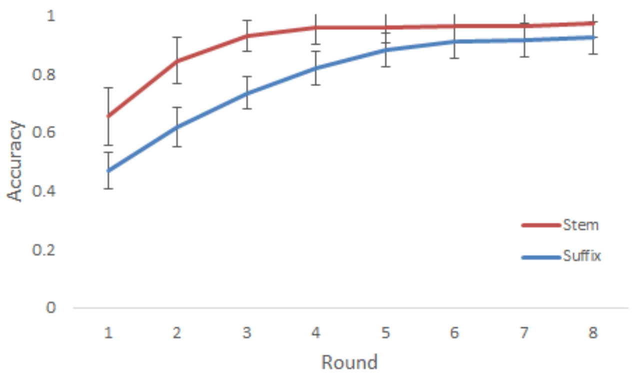


Figure 7: Average stem and suffix accuracy by round. Individual accuracy scores calculated as 1 minus the normalised Levenshtein distance between the production and the target. Error bars are 95% confidence intervals.

one used for the same meaning and word type. Averages by round are illustrated in Figure 8. Suffix stability was significantly greater in Round 8 compared to Round 2 (t(25) = -10.293, p <0.001). Suffix productions are relatively consistent by Round 7 (average suffix stability is 0.91 for Rounds 7 and 8).

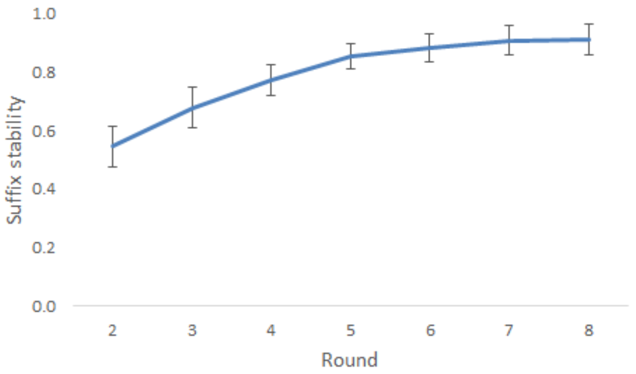


Figure 8: Average suffix stability by round. Individual stability scores calculated as 1 minus the normalised Levenshtein distance between the produced suffix at that round and the previous production for the same stimulus. Error bars are 95% confidence intervals.

3.2.3.2 Measures of complexity

To assess whether adult learners simplify morphology, we compare their suffix productions at Round 2, where they have generally unambiguously acquired the stems of the language but are yet to have acquired the suffixes as accurately, with those of Round 8, where they have acquired or are close to acquiring the target language. At Round 2, average stem accuracy is 0.84 ± 0.08 (95% CI) and average suffix accuracy is 0.62 ± 0.07. At Round 8, average stem accuracy is 0.97 ± 0.05 and average suffix accuracy is 0.92 ± 0.05.[25]

[25] See Section 3.3.1.1 for a repeat of all the complexity measures discussed in this section on a subset of the data which only considers the 12 participants who have a Round 2 stem accuracy score $\geq$ 0.97 (average 0.99 ±

Two example sets of suffixes are shown in Table 6. The stem productions for both these languages exactly matched those of the target language (i.e. stem accuracy = 1). The first language was produced at Round 2. The second is from the same participant's Round 8 productions, and is a case where they have perfectly acquired the target language (suffix accuracy = 1). Comparing both to the meaning space (see Table 5), the Round 2 language is more regular than the Round 8. The quantifier suffixes, for example, are entirely conditioned on Number. In the Round 8 example, they are also dependent on Animal.

	Round 2			Round 8		
	Q	N	V	Q	N	V
1.	a	o	esp	u	o	en
2.	ak	op	esp	uk	op	esp
3.	a	o	esp	u	o	en
4.	ak	op	esp	uk	op	esp
5.	a	o	an	u	o	en
6.	ak	op	an	uk	op	onk
7.	a	o	esp	a	o	an
8.	ak	op	esp	ak	op	asp
9.	a	o	esp	a	o	an
10.	ak	op	esp	ak	op	asp
11.	a	o	an	a	o	an
12.	ak	op	an	ak	op	onk
13.	a	o	esp	a	o	an
14.	ak	op	esp	ak	o	asp
15.	a	o	esp	a	o	an
16.	ak	o	asp	ak	o	asp
17.	a	o	an	a	o	an
18.	ak	op	an	ak	o	onk

Table 6: Example Round 2 and Round 8 suffix sets. Produced by Participant 2. The Round 2 suffix set appears to be "simple" compared to that of Round 8. The Q suffixes are entirely conditioned on Number, with *-a* for singular and *-ak* for plural. The N suffixes are also conditioned on Number, with *-o* for singular and *-op* for plural, bar the exception for Meaning 16. The V suffixes are conditioned on Movement, with V1 and V2 taking *-esp* and V3 taking *-an*, bar the exception for Meaning 16. The Round 8 set is comparatively complex. Both the Q and N suffixes are conditioned on both Number and Animal, while the V suffixes are conditioned on Number, Animal and Movement.

We quantify the complexity of such suffix sets using two different approaches. First we use the established meaning-independent measure of entropy, before we consider a meaning-dependent measure using mutual information.

3.2.3.3 Entropy

The entropy of a set of signals, $H(S)$, is given by:

$$H(S) = -\sum_{s \in S} P(s) \log_2 P(s) \tag{1}$$

0.01).

Entropy is calculated for each of Q, N and V. In our "simple" suffix set of Table 6, for example, we have $H(S_Q) = 1$, $H(S_N) = 0.9911$ and $H(S_V) = 1.1941$. Entropy conditioned on syntactic category, C = {Q,N,V}, can then combine these individual measures:

$$H(S|C) = -\sum_{c \in C} P(c) \sum_{s \in S} P(s|c) \log_2 P(s|c) \tag{2}$$

$H(S|C) = 1.062$ for the "simple" suffix set, compared to the higher conditional entropy of 1.687 for the "complex" example.

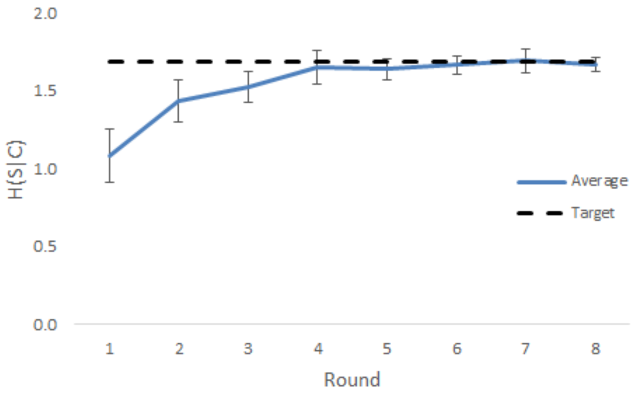


Figure 9: Average $H(S|C)$. Target language conditional entropy = 1.687. Error bars are 95% confidence intervals.

Average conditional entropy over the 8 rounds is illustrated in Figure 9. $H(S|C)$ is significantly lower at Round 2, where the participants are still acquiring the suffixes of the target language, compared to Round 8, where the target language has been (or is close to being) acquired (t(25) = -3.329, p = 0.003). Round 2 entropy is significantly lower than that of the target language (t(25) = -3.871, p <0.001). On the other hand, there is no evidence that entropy at Round 8 differs from that of the target language (t(25) = -0.674, p = 0.507).

This suggests that Round 2 languages are less complex than those of Round 8. We can also consider Q, N and V separately, to see if there is any variation in the complexity of the Round 2 and Round 8 productions across the different word types. $H(S_Q)$, $H(S_N)$ and $H(S_V)$ are show in Figure 10.

$H(S_Q)$ and $H(S_V)$ are significantly lower at Round 2 compared to Round 8 (t(25) = -5.640 and t(25) = -3.343, respectively, p $\leq$ 0.003) and significantly lower than the entropy of the target language ($|t(25)| \geq 3.714$, p $\leq$ 0.001). $H(S_N)$, on the other hand, is significantly greater at Round 2 compared to Round 8 (t(25) = 2.106, p = 0.045). It is also significantly greater than the target language at Round 2 (t(25) = 2.141, p = 0.042). There is no evidence of a difference between any of the measures at Round 8 and the entropy of the target language ($|t(25)| \leq 1.136$, p $\geq$ 0.267).

This meaning-independent measure therefore suggests that the morphological systems produced at Round 2 are generally less complex than those at Round 8. This does not appear to be the case for all three word types, however, as the noun suffixes are more complex in Round 2. This may be due to the noun system being relatively simple (for the target language, $H(S_N) = 0.918$, compared to $H(S_Q) = 1.918$ and $H(S_V) = 2.224$): deviations from the target language

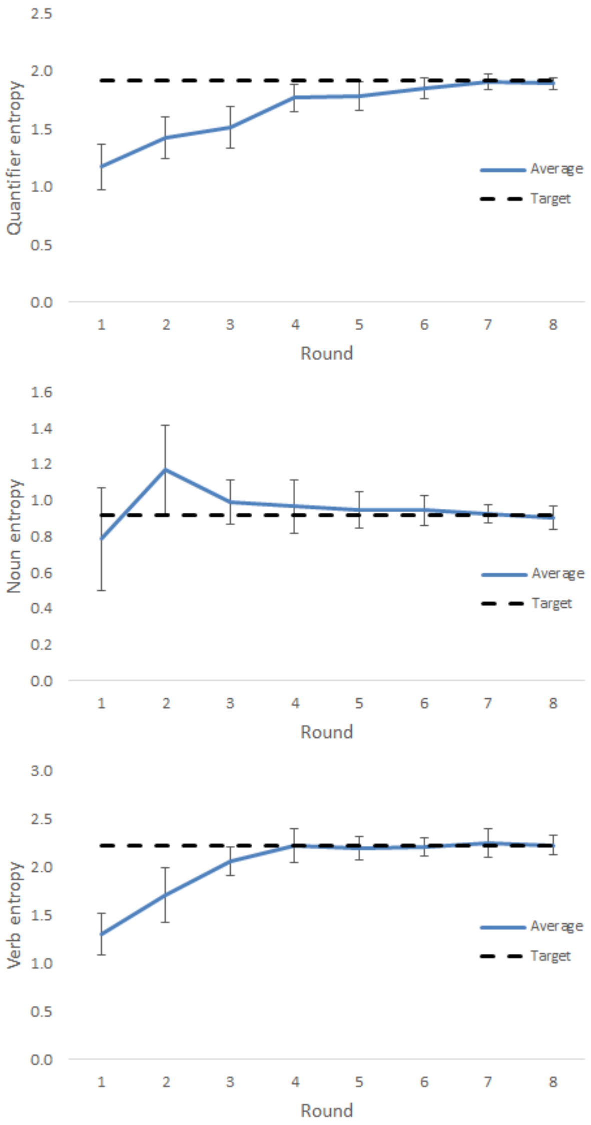


Figure 10: Average $H(S_Q)$, $H(S_N)$, and $H(S_V)$. Target language entropy for Q = 1.918, N = 0.918, and V = 2.224. Error bars are 95% confidence intervals.

noun suffixes may therefore be more likely to increase complexity than for the quantifiers and verbs.

3.2.3.4 Proficiency

We now turn to a meaning-dependent measure of complexity. For a given signal set S and meaning space M, we can quantify their interdependence by calculating mutual information:

$$I(S;M) = \sum_{s\in S}\sum_{m\in M} p(s,m)\log_2 \frac{p(s,m)}{p(s)p(m)} \tag{3}$$

For each of S_Q, S_N and S_V, we consider which dimension of the meaning space (Table 5) is such that the mutual information between the suffix set and the single dimension of the meaning space is maximal. For example, in our “simple” language produced by a participant at Round 2, $I(S_Q; M_{Number}) = 1$, $I(S_Q; M_{Animal}) = 0.739$ and $I(S_Q; M_{Movement}) = 0.059$. So mutual information is greatest when we consider the Number dimension of the meaning space with the quantifier suffixes. For comparison with S_Q of other languages, we create a scalar metric by normalising this maximal value by $H(S_Q)$ (which in this example is 1). This normalisation of mutual information is the proficiency, or uncertainty coefficient, of S given M:

$$C_{MS} = U(S|M) = I(S; M)/H(S) \tag{4}$$

This gives us a measure of the interdependence of a set of suffixes and the most informative single dimension of the meaning space. We term this “Level 1 proficiency”. We can also consider maximal proficiency between the suffix set and pairs of dimensions of the meaning space. Here, we calculate $U(S_Q|M_{Number*Animal}) = 1$, $U(S_Q|M_{Number*Movement}) = 1$ and $U(S_Q|M_{Animal*Movement}) = 1$. The maximum of these values is 1. We term this “Level 2 proficiency”. We can also calculate “Level 3 proficiency” as $U(S_Q|M_{Number*Animal*Movement})$, which, as the mutual information between a suffix set and all three dimensions of the meaning space will be maximally informative, will be equal to 1.[26]

In this case, Level 1 proficiency, Level 2 proficiency and Level 3 proficiency are all equal to 1, indicating that S_Q is comprehensively described by a single dimension of the meaning space, and that knowledge of the other dimensions adds no information: Level 1 = 1, Level 2 increase (= Level 2 proficiency − Level 1 proficiency) = 0, and Level 3 increase (= Level 3 proficiency − Level 2 proficiency) = 0. This is clearly reflected in the Q suffixes in Table 6. For S_N, however, we have Level 1 proficiency = 0.746, Level 2 increase = 0.099 and Level 3 increase = 0.154, suggesting that considering two dimensions of the meaning space is more informative than one, and three more than two. For S_V, Level 1 proficiency = 0.819, Level 2 increase = 0.088 and Level 3 increase = 0.093.

Level 1, Level 2 and Level 3 increases in proficiency for this Round 2 example language of Table 6 are illustrated in Figure 11, alongside our Round 8 example. The Round 2 suffix set appears to be more fully conditioned by a single dimension of the meaning space for all three word types, while proficiency increases by a greater amount by adding dimensions of the meaning space for the Round 8 suffix set.

The average increases in proficiency for all productions at Round 2 and Round 8 are illustrated in Figure 12. For Q, a one-way MANOVA was conducted with *round* (Round 2 or 8) as the independent variable and *Level 1, 2* and *3 increases* as dependent variables. There is a significant effect of *round* (Pillai’s trace = 0.698, F(3,48) = 37.041, p <0.001). Post-hoc multiple comparison tests confirm this effect on each of *Level 1 increase* (t(25) = 2.834, p = 0.009), *Level 2 increase* (t(25) = -8.876, p <0.001), and *Level 3 increase* (t(25) = 4.967, p <0.001).

The noun suffixes were analysed in the same way. There is a significant effect of *round* (Pillai’s trace = 0.169, F(3,48) = 3.261, p = 0.029) on *Level 2 increase* (t(25) = -3.063, p = 0.005). There is no effect on *Level 1* (t(25) = 1.7221, p = 0.097) or *Level 3 increase* (t(25) = 0.383, p= 0.705).

[26] Except in the case where the same suffix is used for all the cases of a word type, when all three measures will be 0. This is only the case for S_V at Round 2 for Participants 21 and 23.

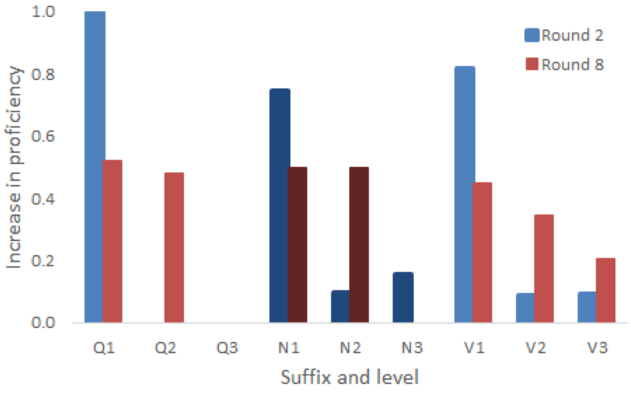


Figure 11: Increases in proficiency at Levels 1, 2 and 3 for example Round 2 and Round 8 languages. The Round 2 suffix set (shown in Table 6) clearly appears to be more conditioned on a single dimension of the meaning space than the Round 8 set (which, for this participant, is the same as the target language).

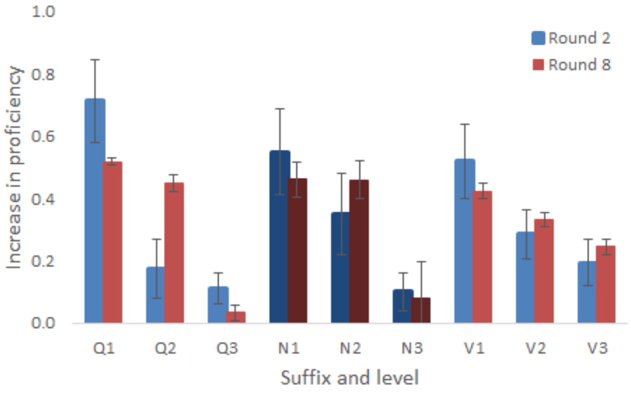


Figure 12: Increases in proficiency. Measures of Q, N and V are shown separately. Error bars are 95% confidence intervals.

For the verb suffixes, there is a significant effect of *round* (Pillai's trace = 0.178, F(3,48) = 3.454, p = 0.024) on *Level 2 increase* (t(25) = -3.511, p = 0.002) and a marginally significant effect on *Level 1* (t(25) = 2.025, p = 0.054), but no effect on *Level 3 increase* (t(25) = -1.638, p = 0.114).

In summary, for each of Q, N and V, Level 2 increase in proficiency is higher for the suffixes of Round 8 productions than for those of Round 2. Q Level 1 proficiency is also significantly higher for Round 2 (and higher with borderline significance in the case of V). So we have evidence here that the earlier round's suffixes are only conditioned on a single dimension of the meaning space, while the later round's are better described by a combination of two meaning space dimensions. The Round 2 morphological systems therefore appear to be less complex.

I also carried out 3 alternative measures of assessing the complexity of the language, calculating statistical complexity (Crutchfield and Young, 1989; Crutchfield and Whalen, 2012), number of rewrite rules, and Mantel tests. This supplementary information can be found in Appendix C.

3.2.4 Discussion and conclusion

The morphological systems of the languages produced by the participants after only 2 rounds of training and testing are simpler than those produced at Round 8. Entropy of the suffix sets at Round 2 is lower than that at Round 8. Proficiency which considers two dimensions of the meaning space is more informative for Round 8 compared to Round 2 for all word types, while one dimension of the meaning space is more informative for the quantifier suffixes for Round 2 compared to Round 8. There is no significant evidence that Level 1 proficiency is greater at Round 2 than at Round 8 for the nouns and verbs, however.

These participants, while attempting to learn a (redundant) morphological system of suffixes, have therefore acquired and produced more simple systems in the earlier stages of their acquisition. This supports the hypothesis that adult learners acquire more regular, and therefore simpler, morphology than that of their target language. In this experiment, with sufficient training and testing, our learners managed to acquire at least a close approximation of the target language. In more naturalistic language learning, however, with a much more challenging target language to acquire, adult learners are likely never to reach native-like competence (Selinker, 1972). If their interlanguage fossilises, and has simpler morphology than the language they are attempting to acquire, then a language with adult learners will have speakers producing simplified morphology which never reaches the level of complexity of that of native speakers. Even if all the non-native speakers do eventually acquire the target language, there will still be a substantial period where they are producing simplified morphology.

As discussed above, this is not sufficient an explanation for why languages with larger proportions of non-native speakers have simpler morphology. All this can explain is how a subset of the population could speak a simpler version of the language.[27] What is also needed is a mechanism by which the simplifications of these non-native learners can affect the language at the level of the population. I consider this question of linkage in Experiments 6 and 7.

3.3 Experiment 6: adult learning and linguistic input

From the languages produced by the Experiment 5 participants at Round 2, we have a naturalistic set of morphological systems which are simpler than that of the learners' target language. By Round 8, however, the produced suffix sets were typically very similar if not identical to those of the target language (21 of the 26 participants having a suffix success score greater than 0.90), and so the languages exhibit a comparable degree of morphological complexity. Though such close-to-ceiling performance is unlikely through adult learning of a real language (Selinker, 1972), this has been possible in this miniature language, with the minor differences between speakers in their Round 8 productions representing some natural-like variation. Experiment 6 uses these outputs from Experiment 5 as inputs for a second set of learners. Participants either receive "Complex input", in which their training data is provided solely by Experiment 5's complex, Round 8 languages, or "Mixed input", in which they receive a combination of data from the complex and simple languages. We can then investigate how the simple languages in the input can affect the complexity of the languages acquired by our second generation of participants. The construction of these input sets from the Experiment 5 productions is illustrated in Figure 13.

[27] Unless the complexity of the language as a whole were defined as some average measure of the complexity of the individual languages of all of its speakers.

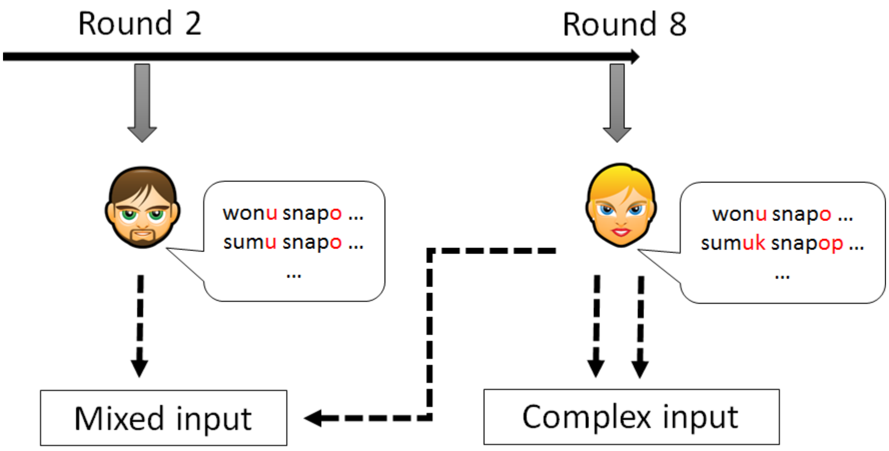


Figure 13: Construction of input data for Experiment 6 from Experiment 5's productions. "Non-native" languages are taken from Experiment 5's productions at Round 2 and "native" from Round 8. Linguistic input comprised only of native languages is then "complex" input, while "mixed" input is a combination of native and non-native languages.

In this experiment, I investigate how adult-learner simplifications, such as those produced in Round 2 of Experiment 5, are affected by transmission, through the language learning of the second generation of learners. In natural-language learning, linguistic input will be made up from the productions of the learner's speech community. If that speech community is large, the input may be provided by a greater number of speakers (Hay and Bauer, 2007; Nettle, 2012). If that speech community has a higher proportion of non-native speakers, the input may include more non-native speaker productions, which may include simpler morphology than that produced by the native speakers. Of interest is how the languages which are acquired only from native speaker productions compare with those acquired from a mix of native speaker and non-native speaker productions. If the languages learned in the second case have simpler morphology than those in the first, then language simplification at group level could be solely explained by language learning: simplifications at the level of the individual adult learner could diffuse by forming part of the linguistic input for future learners.

To determine how simplified linguistic input may affect language learning when mixed with complex, the second set of participants attempt to acquire a language from linguistic input taken from the productions of multiple participants from Experiment 5. For a given participant, this was dependent on two variables. The first was the proportion of speakers who provide the input who produce simplified morphological systems, to see if simplifications in the input results in simpler languages being acquired. The second was group size, which was included to see if there is any interaction between the proportion of the input which was output from simplified language systems, and the total number of speakers who provide the input. This relates to the proposal that it is the proportion of non-native speakers, rather than number of speakers, which is the direct determinant of linguistic complexity. We may predict that the proportion of simplified input will result in the acquisition of simpler morphological systems, but that group size will have no effect.

Given the size of the miniature languages which the learners acquire and so the potential to achieve as complex a language as that of the input for our adult participants, this experiment can shed some light on how simplifications arising from adult learning may affect the languages acquired by child learners in natural language learning.

3.3.1 Materials and methods

The participants who had best acquired the stems in Round 2 were identified; those whose stem productions *exactly* matched the stems of the target language for at least 90% of the labels. 12 participants met this criteria, and had an average Round 2 stem success score of 0.99. The Round 2 stem and suffix productions of these participants then formed a set of simplified "non-native" input data, here produced by adult learners of a language. The Round 8 data for these participants, with average stem success of 1.00 and average suffix success of 0.98, formed a set of complex input data. Given its proximity to the target language of Experiment 5, we can consider this a proxy for input data from "native" speakers of the language, with only some comparatively minor variations in production amongst speakers.[28]

3.3.1.1 Analysis of input data set

I repeated the analyses of Experiment 5 on this subset of the data to confirm that the morphological systems displayed similar characteristics to those of the larger set.

Stem (t(11) = -3.304, p = 0.007) and suffix accuracy (t(11) = -13.347, p <0.001) are significantly greater at Round 8 compared to Round 1. Suffix stability is significantly greater at Round 8 compared to Round 2 (t(11) = - 10.073, p <0.001).

Entropy of the suffix sets at Round 2 is significantly lower than at Round 8 (t(12.572) = 4.337, p <0.001). Entropy at Round 2 is also significantly lower than that of the target language (t(11) = -4.169, p = 0.002), while there is no evidence of a difference between Round 8 and target language entropies (t(11) = 1.197, p = 0.256).

For the proficiency measure, there is still a significant difference between Round 2 and Round 8 productions for Q (Pillai's trace = 0.783, F(3,20) = 24.044), p <0.001). *Level 1 proficiency* was higher for Round 2 than Round 8, (t(11) = 3.354, p = 0.006). *Level 2 increase* was higher for Round 8 than Round 2 (t(11) = -6.587, p <0.001). Unlike in the full data for all 26 participants in Experiment 5, *Level 3 increase* was greater at Round 2 than at Round 8, (t(11) = 3.675, p = 0.004).

There were no similar significant effects for the noun suffixes (Pillai's trace = 0.128, F(3,22) = 0.977, p = 0.423) or the verb suffixes (Pillai's trace = 0.141, F(3,22) = 0.141, p = 0.378).

This subset of the Experiment 5 data appears to be generally representative of the larger set, and at least qualitatively demonstrates the same trend of an increase in complexity from the Round 2 data to the Round 8. The reduction of the sample size will have reduced statistical power, however, and this likely explains the non-significant differences between Round 2 and Round 8 for the noun and verb suffixes under the proficiency measure.

3.3.1.2 Procedure

The general design follows that of the previous experiment. Participants were required to learn

[28] Though it would have been possible to use all the Round 2 and 8 Experiment 5 productions as input for Experiment 6, I limited the set so as to avoid there being any significant variation in the stems the learners received as input.

a language which described the same image set and meaning space (see Figure 6 and Table 5) over 8 rounds, with the same procedure for training and testing described in Section 3.2.1.

The input data, however, was dependent on condition. Learning in four different types of social group was considered:

- A *small* population (n=2) in which *all* speakers were native speakers (Small-AllNative);
- A *large* population (n=8) in which *all* speakers were native speakers (Large-AllNative);
- A *small* population (n=2) in which *half* the speakers were native speakers and half were non-native (Small-HalfNative);
- A *large* population (n=8) in which *half* the speakers were native speakers and half were non-native (Large-HalfNative).

12 participants were randomly assigned to each of the 4 conditions, and randomly assigned model input speakers accordingly. In the Small-AllNative condition, the data of 2 "native" speakers were selected. In the Large-AllNative condition, 8 sets of "native" speaker data were selected. In the Small-HalfNative, the data of 1 "non-native" speaker and 1 "native" were selected, with the condition that the two sets were not originally produced by the same participant of Experiment 5. In the Large-HalfNative condition, 4 sets of "non-native" data and 4 sets of "native" data were selected, again with the stipulation that no Experiment 5 participant produced both "native" and "non-native" input data.

In a training round, 9 images were randomly selected for input. These were then equally divided (as far as possible, given that the size of the group was either 2 or 8) and randomly allocated to the input speakers. The input speaker's corresponding label was then presented alongside the image. These training items were then presented as in Experiment 5, with no overt marking of speaker identity. As the images and speakers were randomly selected for each training round, a participant could potentially see a different label with the same image from one round to the next.

Evidence of the Small-HalfNative and Large-HalfNative languages being less complex than Small-AllNative and Large-AllNative would support the proposal that non-native speakers can contribute to language-level morphological simplification. The individual-level simplifications would have reduced the complexity of the language acquired by the subsequent generation.

3.3.2 Participants

48 native English speakers (21 male; aged between 19 and 40, mean 23.2) were recruited at the University of Edinburgh. Each was compensated £7. The experiment was run between 16th June 2014 and 12th February 2015.[29]

3.3.3 Analysis and results

I consider the stem and suffix acquisition, and the complexity of the participant productions using the same measures as in Experiment 5. The key difference is that we are now interested

[29] The data for this experiment originally involved 20 participants in each condition (n=80) and was collected between 16th June and 14th July 2014. Due to a programming error, not all the participants were assigned input speakers randomly as intended, and so data for these 37 participants was rejected and is not included in any analysis here. Data for the final 6 participants who make up this sample of 48 was collected on 12th February 2015. Analysis of the original data instead leads to the same overall results and conclusions described here.

in a comparison of the final (Round 8) productions of each language across the conditions, as opposed to Experiment 5's comparison of Round 2 and Round 8 productions.

3.3.3.1 Acquisition of stems and suffixes

As the input data varies between and within conditions, repeating the stem and suffix acquisition analysis of Experiment 5 does not evaluate learning success for each participant in the same way. As the speakers who provide an individual's input may also not have identical labels for a given image, a given participant may also not have been receiving consistent input. Despite this, comparing the acquisitions of the stems and suffixes to the target language of Experiment 5 may still be useful in highlighting any key similarities and differences between the conditions. Stem accuracy, suffix accuracy, and change in suffix between rounds were therefore calculated as in Experiment 5, and Figures 14, 15, and 16 show the average for each measure for each condition.

A linear mixed effects analysis for stem accuracy included *population size* (small or large), *population type* (all native or 50% non-native), *round* and their interactions as (centred) fixed effects. *Participant identity* was investigated as a random effect. This model was significantly better than the equivalent null model ($\chi^2(7) = 157.07$, p <0.001). Analysis for suffix accuracy and suffix stability were carried out in the same way. Each model was again significantly better than its equivalent null model ($\chi^2(7) = 342.71$, p <0.001, and $\chi^2(7) = 231.92$, p <0.001, respectively). The results of the models are summarised in Table 7.

Measure	Fixed effect	β	SE	t	Pr(>| t|)	
Stem accuracy	population size	-0.012	0.024	-0.49	0.624	
	population type	-0.047	0.024	-1.97	0.050	*
	round	0.042	0.003	13.57	<0.001	***
	pop.size*pop.type	-0.063	0.048	-1.33	0.184	
	pop.size*round	-0.003	0.006	-0.44	0.660	
	pop.type*round	0.009	0.006	1.51	0.132	
	pop.size*pop.type*round	0.017	0.012	1.36	0.175	
Suffix accuracy	population size	0.019	0.035	0.54	0.590	
	population type	-0.120	0.035	-3.48	<0.001	***
	round	0.054	0.002	23.40	<0.001	***
	pop.size*pop.type	-0.046	0.069	-0.67	0.503	
	pop.size*round	-0.005	0.005	-1.01	0.313	
	pop.type*round	-0.009	0.005	-1.97	0.050	*
	pop.size*pop.type*round	0.016	0.009	1.71	0.089	.
Suffix stability	population size	-0.020	0.029	-0.68	0.496	
	population type	0.082	0.029	2.82	0.005	**
	round	-0.058	0.003	-18.13	<0.001	***
	pop.size*pop.type	0.008	0.058	0.14	0.888	
	pop.size*round	-0.005	0.006	-0.74	0.459	
	pop.type*round	0.008	0.006	1.18	0.238	
	pop.size*pop.type*round	-0.008	0.013	-0.64	0.522	

Table 7: Summary of linear mixed modelling results for stem and suffix acquisition measures.

Stem accuracy with respect to the target language of Experiment 5 therefore increases with round, with Small-AllNative and Large-AllNative conditions being marginally more accurate than Small-HalfNative and Large-HalfNative. The final production stem accuracy scores are

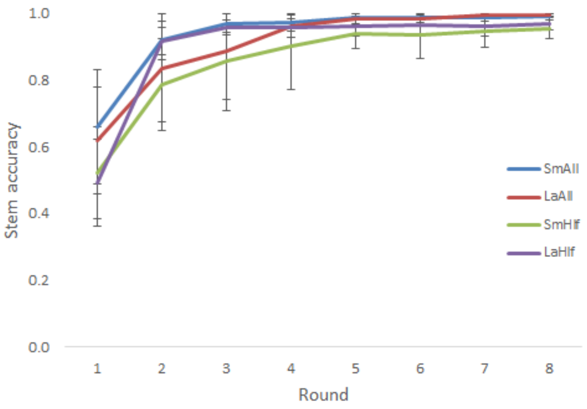


Figure 14: Average stem accuracy by round and condition. Individual accuracy scores calculated as 1 minus the normalised Levenshtein distance between the production and the target of Experiment 5. Error bars are 95% confidence intervals.

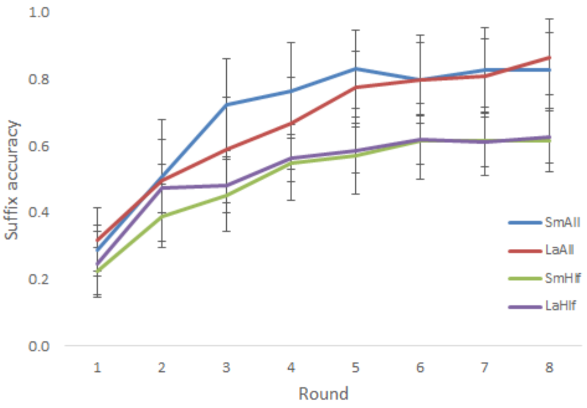


Figure 15: Average suffix accuracy by round and condition. Individual accuracy scores calculated as 1 minus the normalised Levenshtein distance between the production and the target of Experiment 5. Error bars are 95% confidence intervals.

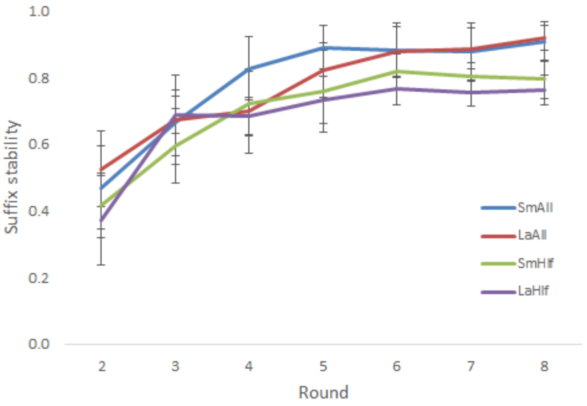


Figure 16: Average suffix stability by round and condition. Individual accuracy scores calculated as 1 minus the normalised Levenshtein distance between the produced suffix at the at round and the previous production for the same stimulus. Error bars are 95% confidence intervals.

also very high (only 1 of the 48 participants had a score less than 0.9), suggesting that there is little variation from the original target language stems across all conditions. Suffix accuracy also increases with round, with the Small-AllNative and Large-AllNative languages reproducing the target language of Experiment 5 significantly more accurately than the Small-HalfNative and Large-HalfNative.

3.3.3.2 Entropy

Average entropy scores by condition are illustrated in Figure 17. Considering Q, N and V as separate dependent variables, there is no significant difference between the conditions in the final round (Pillai's trace = 0.264, F(9,132) = 1.417, p = 0.187).

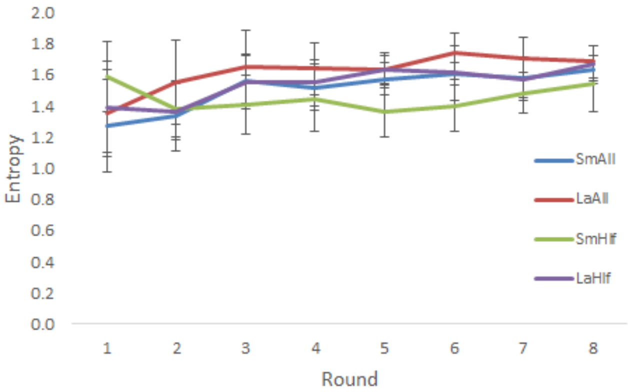


Figure 17: Average combined entropy by round and condition. Error bars are 95% confidence intervals.

3.3.3.3 Proficiency

Figure 18 shows the differences between conditions for the increases in proficiency for Q, N and V.[30] Figure 19 gives the averages across the word types, for illustration only.

There was no significant difference between conditions for the quantifier (Pillai's trace = 0.295, F(9,132) = 1.598, p = 0.122) and noun suffixes (Pillai's trace = 0.250, F(9,132) = 1.334, p = 0.225). There was a condition-dependent difference for the verb suffix (Pillai's trace = 0.433, F(9,132) = 2.476, p = 0.012), however: for *Level 1 increase* (F(3,44) = 4.191, p = 0.011), Small-HalfNative was significantly higher than Large-HalfNative (Tukey's HSD, p = 0.008). No other differences between Level 1 means were significant ($p \geq 0.063$). There was also a significant difference between conditions for *Level 3 increase* (F(3,44) = 3.727, p = 0.012), with Large-HalfNative greater than Small-HalfNative (p = 0.011), but no differences between any of the other Level 3 increase means ($p \geq 0.167$). There was no difference for *Level 2 increase* (F(3,44) = 0.971, p = 0.415).

There is generally no evidence of differences between the conditions. For the verb suffixes, however, there is some indication that Small-HalfNative may be simpler than Large-HalfNative. This would be in contrast to the claim that languages spoken by more people have less complex morphological systems (Lupyan and Dale, 2010).

[30] As in Experiment 5, this will sum to 1 across the 3 levels for each of Q, N and V for a given participant, as the mutual information of a suffix set and all three dimensions of the meaning space will be maximally informative. The exception to this is where the same suffix is applied to every occurrence of a word type. This is only the case for the noun set of a single participant.

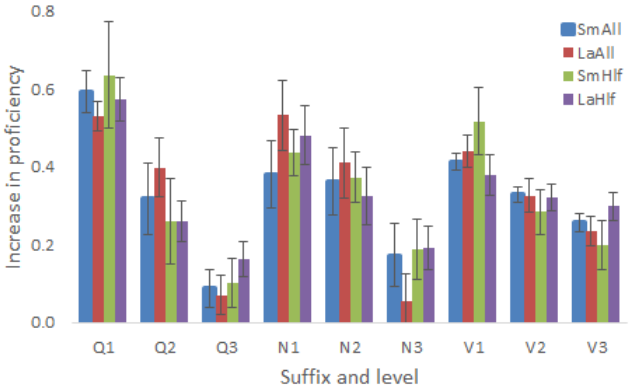


Figure 18: Average increases in proficiency by level. Measures of Q, N and V are shown separately. Error bars are 95% confidence intervals.

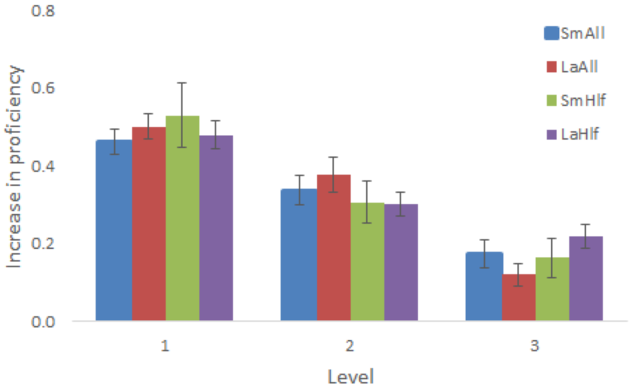


Figure 19: Average increases in proficiency by level. These averages across word types are shown to illustrate any broad differences between the conditions; analysis considers each word type separately. Error bars are 95% confidence intervals.

As for Experiment 5, I also carried out 3 alternative measures of assessing the complexity of the language, calculating statistical complexity (Crutchfield and Young, 1989; Crutchfield and Whalen, 2012), number of rewrite rules and Mantel tests. This supplementary information can be found in Appendix D, Sections D.1 to D.3.

3.3.4 Interim discussion

There is very limited evidence of a condition-dependent difference in the complexity of the acquired morphological systems. Where there is some evidence of an effect of condition on the complexity of verbal morphology, it indicates that languages with greater numbers of speakers are more complex, in contrast to the claim that languages with lower numbers of speakers are the more complex (Lupyan and Dale, 2010).

Considering the stem and suffix success measures (which do not directly quantify system complexity), there are indications of the two groups with only native speaker inputs more accurately acquiring the original target language from Experiment 5. The languages acquired in the two groups consisting of a mix of native and non-native speakers have diverged from the

Experiment 5 target language to a greater extent. The similarity between the two all native groups and between the two half non-native conditions would suggest that of the two proposed determinants of morphological complexity, proportion of non-native speakers is a more likely determinant than number of speakers.

In the Small-HalfNative and Large-HalfNative conditions, the participants received a mix of comparatively complex and simple morphological systems. Compared to the Small-AllNative and Large-AllNative participants who were only exposed to complex morphological data, it may be surprising that they did not acquire simpler systems. The following section aims to try and explain this result, and also shed some more light on the differences between the produced languages in the HalfNative and AllNative conditions. In Section 3.3.5, I take a closer look at the input the participants received in each condition. In Section 3.3.6, I then focus on the suffix forms they produced.

3.3.5 Input analysis

I consider three different ways of analysing the input: a simple entropy measure analogous to the output entropy measure, a meaning-conditioned entropy measure, and proficiency.

3.3.5.1 Entropy

The (meaning independent) entropy was calculated as for the output data for each word type, but for the full input data set of 72 items for each word type per participant, as opposed to the 18 items of their output. Comparison by word type and condition is shown in Figure 20.

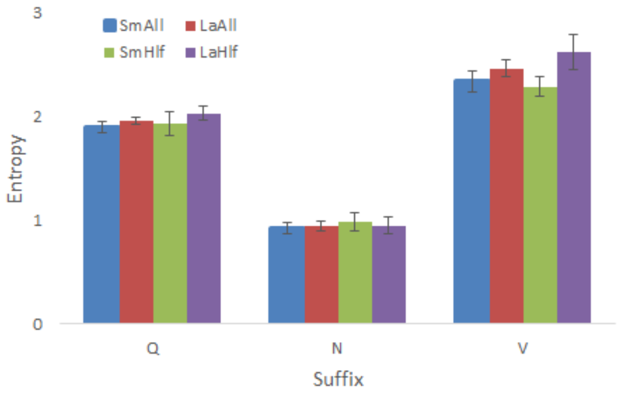


Figure 20: Input variability entropy. Error bars are 95% confidence intervals.

There was a significant effect of *population size* (Pillai's trace = 0.332, F(3,42) = 6.971, p <0.001) on *quantifier suffix entropy* (F(1,44) = 5.713, p = 0.021) and *verb suffix entropy* (F(1,44) = 18.656, p <0.001), but not on *noun suffix entropy* (F(1,44) = 0.0833, p = 0.774). There was no effect of *population type* (Pillai's trace = 0.077, F(3,42) = 1.171, p = 0.332) or its interaction with *population size* (Pillai's trace = 0.101, F(3,42) = 1.565, p = 0.212).

This would imply that the only differences in the variability of the inputs is due to the number of input speakers, with more variability in the quantifier and verb suffixes in the larger populations.

3.3.5.2 Entropy conditioned on meanings

We can also consider a meaning-dependent measure of entropy. Here the entropy is first calculated for the inputs of each word for each meaning. This is then averaged across all of the meanings for a given word type. The average of these measures by word type and condition is illustrated in Figure 21.[31]

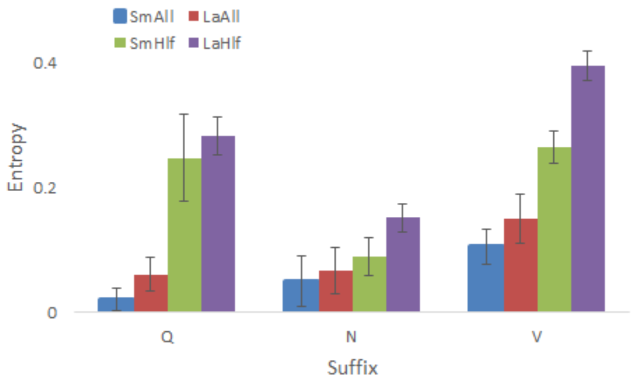


Figure 21: Input variability entropy conditioned on meaning. Error bars are 95% confidence intervals.

There was a significant effect of *population size* (Pillai's trace = 0.527, F(3,42) = 15.614, p <0.001) on the entropy of the suffixes for all syntactic categories (F(1,44) $\geq$ 4.062, p $\leq$ 0.050). There was also a significant effect of *population type* (Pillai's trace = 0.898, F(3,42) = 123.458, p <0.001) for all categories (F(1,44) $\geq$ 16.571, p <0.001). There was a significant effect of the interaction of *population size* and *type* (Pillai's trace = 0.202, F(3,42) = 3.543, p = 0.022) on *verb suffix entropy* (F(1,44) = 10.339, p = 0.002), but not on *quantifier* or *noun suffix entropy* (F(1,44) $\leq$ 2.353, p $\geq$ 0.132).

Here we can see more striking differences between the conditions: large populations produce more variable input for all syntactic categories, as do those with non-native speakers. There is also an interaction effect in the verb suffixes, with large mixed populations resulting in more variable input.

As an example of this effect, we can consider 4 example participants, 1 from each condition, and look at the data they received for the Quantifier suffix for the "1 crocodile straight" image over the 9 rounds of training. The Small-AllNative participant was exposed to the suffix *-u* 5 times, as was the Large-AllNative participant. The Small-HalfNative participant was exposed to *-u* 3 times and *-a* twice. The Large-HalfNative participant was exposed to *-u* 3 times and *-o* 3 times. For this particular image and suffix, there is no variation in the data for the Small-AllNative and Large-AllNative participants ($H(S) = 0$), but there is for the Small-HalfNative ($H(S)$ =0.97) and the Large-HalfNative ($H(S)$ =1). This is largely intuitive: input is likely to be more variable when it comes from groups of speakers with mixed levels of linguistic competence.

[31] This measure differs from conditional entropy, $H(S|M)$, as 9 stimulus-label pairings were randomly-selected for each training round. Therefore a participant did not necessarily receive all images exactly the same number of times.

3.3.5.3 Proficiency

Finally, I consider the proficiency between the input data and the meaning space. This is calculated as for the output data in Experiments 5 and 6, but with a necessary adjustment. In the analysis of the output data, the *Level 3 proficiency*, which we defined as the mutual information between all 3 variables of the meaning space and the suffix for a given word type, will necessarily be 1: knowing exactly what the image is will enable you to identify the suffix used. Therefore we considered *Level 1 proficiency*, *Level 2 increase* (the increase in proficiency by considering 2, rather than only 1, dimensions of the meaning space), and *Level 3 increase*, which would necessarily sum to 1.[32] This is not the case here where we consider the input data, as any variability in the suffixes for a particular image and syntactic category (discussed in Section 3.3.5.2) would lead to *Level 3 proficiency* being less than 1: knowing exactly what the image is would not necessarily enable you to identify the suffix, as the form of the suffix is not consistent.

Therefore instead of considering the increases in proficiency as before, I simply calculate the proficiency between the set of suffixes for a given word type and the relevant number of dimensions of the meaning space, rather than considering the increases at each level. The average proficiency at each level (averaged across the word types for illustration) is shown in Figure 22.

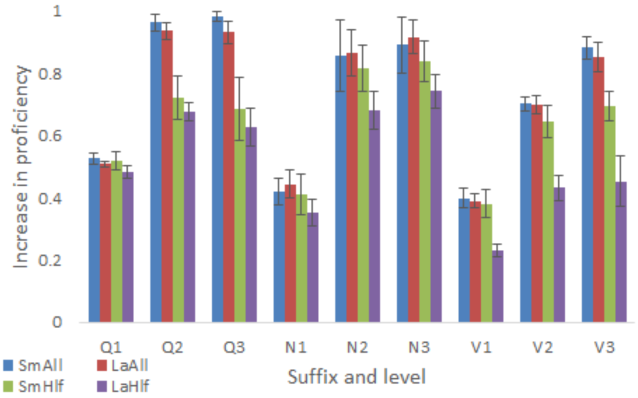


Figure 22: Input proficiency by level. Q, N, and V are shown separately. Error bars are 95% confidence intervals.

Each of the quantifier, noun and verb suffixes were considered separately to assess the effect of *condition* on *Level 1*, *2* and *3 proficiency* in each case.

There was a significant effect of condition for the quantifier suffixes (Pillai's trace = 0.987, F(9,132) = 7.188, p <0.001), for each of *Level 1* (F(3,44) = 4.217, p = 0.010), *Level 2* (F(3,44) = 55.577, p <0.001) and *Level 3 proficiency* (F(3,44) = 38.564, <0.001). For *Level 1 proficiency*, Small-AllNative is significantly greater than Large-HalfNative (diff = 0.044, p = 0.011), and Small-HalfNative greater than Large-HalfNative (diff = 0.038, p = 0.033). No other means were significantly different ($p \geq 0.243$). For *Level 2 proficiency*, Large-AllNative was significantly greater than Large-HalfNative (diff = 0.260, p <0.001), Large-AllNative greater than Small-HalfNative (diff = 0.215, p <0.001), Small-AllNative greater than Large-HalfNative (diff = 0.286, p <0.001) and Small-AllNative greater than Small-HalfNative (diff = 0.241, p

[32] As previously noted, except in the rare cases where the same suffix is applied to all cases for a given word type, where each of these measures will be 0.

<0.001). The other differences were non-significant (p $\geq$ 0.376). For *Level 3 mutual information*, Large-AllNative is significantly greater than Large-HalfNative (diff = 0.307, p <0.001), Large-AllNative greater than Small-HalfNative (diff = 0.245, p <0.001), Small-AllNative greater than Large-HalfNative (diff = 0.357, p <0.001), and Small-AllNative greater than Small-HalfNative (diff = 0.300, p <0.001). The other differences were non-significant (p $\geq$ 0.428).

For the noun suffixes, *condition* had no significant effect (Pillai's trace = 0.350, F(9,132) = 1.937, p= 0.052).

For the verb suffixes, there was a significant effect of condition (Pillai's trace = 1.054, F(9,132) = 7.941, p <0.001) and for each of *Level 1* (F(3,44) = 31.665, p <0.001), *Level 2* (F(3,44) = 56.307, p <0.001) and *Level 3 proficiency* (F(3,44) = 61.86, p <0.001). For *Level 1*, Large-AllNative was significantly greater than Large-HalfNative (diff = 0.159, p <0.001), Small-AllNative greater than Large-HalfNative (diff = 0.168, p <0.001), and Small-HalfNative greater than Large-HalfNative (diff = 0.150, p <0.001). The other differences were non-significant (p $\geq$ 0.800). For *Level 2*, Large-AllNative was significantly greater than Large-HalfNative (diff = 0.267, p <0.001), Small-AllNative greater than Large-HalfNative (diff = 0.271, p <0.001) and Small-HalfNative greater than Large-HalfNative (diff = 0.215, p <0.001). The other differences were non-significant (p $\geq$ 0.105). For *Level 3*, Large-AllNative was significantly greater than Large-HalfNative (diff = 0.401, p <0.001), Large-AllNative greater than Small-HalfNative (diff = 0.156, p <0.001), Small-AllNative greater than Large-HalfNative (diff = 0.431, p <0.001), Small-HalfNative greater than Large-HalfNative (diff = 0.244, p <0.001), and Small-AllNative greater than Small-AllNative (diff = 0.187, p <0.001). There was no significant difference between Small-AllNative and Large-AllNative (p = 0.822).

Overall, proficiency between the suffixes and the meaning space is greater for the input of the AllNative conditions compared to the HalfNative conditions, and greater for Small-HalfNative than for Large-HalfNative. This would suggest that the input data is both more informative about and predictable from the meaning space in the Small and the AllNative conditions, suggesting lower variability. This is intuitive: input from a greater number of speakers is likely to be more variable, as is input from groups of speakers with mixed levels of linguistic competence.

3.3.6 Suffix origins

We have seen little difference in the complexity of the morphological systems that the participants are producing in different groups, yet significant differences in the variation they have received. If we more directly compare their produced suffixes with their inputs, we can get more of an insight into which suffix forms they are using, where they acquired them from, and which elements of the acquisition they are finding more challenging.

First, we consider a variable of *Round Match*, and see how often the participants produce the suffix that they have just seen in that round's training for a given image. We only consider the 9 images seen in the training for a particular round and give a score of 1 if the output label matches the input for each word type and image, and 0 otherwise. The average scores by condition are illustrated in Figure 23.

For the Round 8 testing data, a linear mixed model with Round Match as dependent variable and with *population size*, *type*, and their interaction as fixed effects and *participant identity* as a random effect was significantly better than the null model ($\chi^2(3)$ = 18.497, p <0.001). There

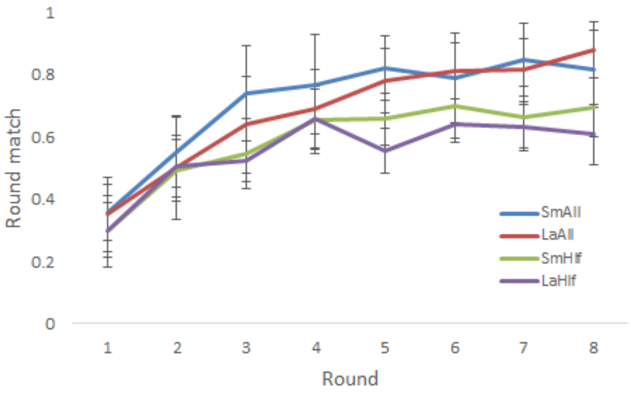


Figure 23: Proportion of suffixes which match the input of that round. Error bars are 95% confidence intervals.

was a significant effect of *population type* (β = -0.272, SE = 0.066, t(141) = -4.134, p <0.001), but not of the other fixed effects ($|\beta| \leq 0.015$, SE $\geq$ 0.066, $|t(141)| \leq 1.595$, p $\geq$ 0.113).

There is therefore evidence that by the final round of training and testing, participants in the Small-AllNative and Large-AllNative more accurately produced the suffixes they saw in Round 8 training. As the input data is more variable in the Small-HalfNative and Large-HalfNative conditions, however, the participants in all conditions may still be producing the majority variant of each suffix. We can quantify this with *Round Majority*: the proportion of suffixes which exactly match the most-observed suffix seen for the same meanings in all the training input up to this point. We consider all of the suffixes for which the corresponding image has been observed at least once in training, and this is illustrated in Figure 24.

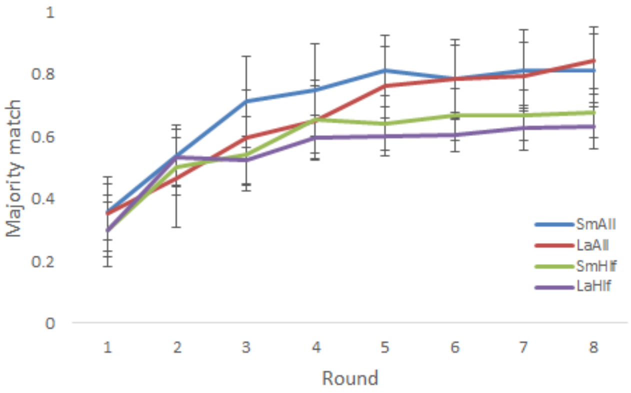


Figure 24: Proportion of suffixes which match the majority input of all rounds up to that point in the experiment. Error bars are 95% confidence intervals.

For the Round 8 testing data, a linear mixed model with Round Majority as dependent variable and with *population size*, *type*, and their interaction as fixed effects and *participant identity* as a random effect was significantly better than the null model ($\chi^2(3)$ = 15.142, p = 0.002). There was a significant effect of *population type* (β = -0.208, SE = 0.062, t(141) = -3.372, p = 0.001), but not of the other fixed effects ($|\beta| \leq 0.071$, SE $\geq$ 0.062, $|t(141)| \leq 0.815$,

$p \geq 0.416$).

It appears that the participants in the Small-AllNative and Large-AllNative conditions are using the majority form they have been trained on for each image more than the Small-HalfNative and Large-HalfNative participants. We can also investigate whether the participants are producing suffix forms they have seen somewhere in training, even if the form-meaning pairing of the training are not matching that of the produced suffixes, by considering *Anywhere Present*, the proportion of suffixes which exactly match one of those observed in *any* of the input received up to that point, illustrated in Figure 25. For this measure, all 18 produced stems for each round are included. Here, the mixed model was no better than the null ($\chi^2(3) = 2.249$, $p = 0.522$), and so there is no suggestion that participants in different *population sizes* or *types* are producing suffixes which they had not been exposed to more than another.[33]

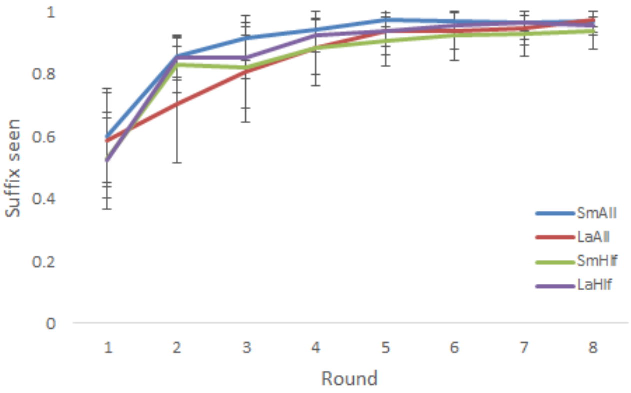


Figure 25: Proportion of suffixes which match one observed anywhere so far. Error bars are 95% confidence intervals.

It would appear then that the AllNative condition learners are more accurately reproducing the suffixes observed for the same meanings, whether only considering the Round 8 training data or the majority suffixes observed from all the training data.[34] The HlfNative condition learners are significantly less accurate for both measures, however, and the relatively low scores (averaging 0.65 for Round Match and 0.65 for Majority Match at Round 8) suggest that they are generally having some difficulty mapping their training input suffixes to meanings. The very limited evidence of differences between the conditions Round Anywhere measure, however, suggest that the HlfNative learners are not having trouble acquiring the forms of the suffixes as accurately as the AllNative. It is mapping the forms to meanings where they are comparatively underperforming.

From Figures 23 and 24, the differences between the AllNative and HlfNative conditions are only clearly apparent in the second half of the experiment. If we consider the data at Round 2, for example, there is no evidence of any difference between the conditions for any of the suffix types for Round Match or Majority Match (comparison with null models: $\chi^2(3) \leq 2.238$, $p \geq 0.525$). This would suggest that with limited training and testing, both participants which received half non-native input and those who received only native input were managing to acquire the suffix forms if not map them to meaning equally well, but while the lack of variability in the

[33] An alternative measure of *Round Present*, the proportion of suffixes which exactly match any of those observed in that particular round's training, gives the same result ($\chi^2(3) = 6.459$, $p = 0.091$).

[34] In the AllNative conditions, due to the limited variability in the input, these measures are very similar.

no non-native input conditions allowed acquisition of the form-meaning mappings, increased variability prevented this from happening in the other conditions.

3.3.7 Discussion and conclusion

The lack of difference between our conditions offers no support for the hypothesis that simply receiving input data from a mix of simplified and complex data is enough for morphological simplifications arising from adult learning to reduce the complexity of a language at the level of the group.

We have seen that adult learners produce simpler morphology than that of their target language (Experiment 5), that mixing this comparatively simple non-native language with native language results in high variability for a next generation of learners (Section 3.3.5), and that the effect of receiving this input is the acquisition of morphological systems which are no more simple than that acquired by learners who only receive their input from native speakers (Section 3.3.3). This appears to be due to the participants who received the mixed input from native and non-native speakers being unable to track the form-meaning mappings for the suffixes (Section 3.3.6).

If we return to the data of Experiment 5, we have can also see evidence of adult learners being unable to track the form-meaning mappings for the suffixes. Calculating Round Match and Round Present (from Section 3.3.6) for the Round 2 data of Experiment 5 (from which we extracted our "non-native" input data for this experiment), we can again see evidence of a failure to acquire form-meaning mappings: Round Match score is only 49% ± 0.08 (95% confidence intervals) and Round Present only 77% ± 0.08.[35] It would appear that the relative simplicity of the suffix morphological systems at Round 2 does not solely reflect the participants acquiring the form-meaning mappings of the training data and generalising to novel images. There also appears to be some loss of the mappings to meanings for the suffixes seen in training.

The relevance of this is that our participants in the Small-HalfNative and Large-HalfNative conditions are not receiving a mix of complex, "native" input and simplified variants of that input: sets of labels which are constructed by generalising a subset of the complex input to the whole meaning space. Rather, they are receiving a combination of a complex morphological system and a simpler morphological system, but where the simpler system may have little relationship to the complex. To illustrate this, consider our original target language's (Table 5 on p. 56) noun suffixes: the suffix *-o* is used to mark singular nouns and *-op* to mark plural, except for the "bird" noun class, with takes the *-o* suffixes regardless of number. A simplified, *generalised-subset* of this would be consistent use of *-o* for singular and *-op* for plural for all nouns, or simpler *-o* for all nouns regardless of number. But what our participants are receiving is not necessarily simplified *generalised-subsets* of the target language, but other more simplified systems. They could, in an extreme example, receive *-a* for all nouns and numbers, or *-op* for all birds and *-u* otherwise. These are simpler systems, but they are not versions of the target language with reduced complexity.

We can have a look within our data, though, to find examples where the participants *have* been receiving a mix of complex data and a generalised subset of this data. Looking at the input

[35] Both measures are significantly greater at Round 8 compared to Round 2 (t(49.28) = -8.341 and t(32.22) = -4.636, respectively, p <0.001). The high Round 8 scores (Round Match = 0.91 ± 0.07, 95% CI; Round Present = 0.97 ± 0.03) is to be expected given their very accurate acquisition of the target language.

data for the Small-HalfNative participants,[36] I categorised the non-native half of the input as being either *Generalised-subset* or *Non-generalised-subset*. A Generalised-subset input would be a simpler and simplified variant of the complex input for that participant. A Non-generalised-subset would be simpler but not a generalised-subset of the complex input. Categorisation was done by hand, and for both the quantifier and noun suffixes, which were considered separately.[37] For the quantifiers, 6 participants were categorised as having received Generalised-subset input, and 5 as Non-generalised-subset. For the nouns, 5 were categorised as Generalised-subset, and 6 Non-generalised-subset. The output entropy for this subset of the main data is shown in Figure 26.

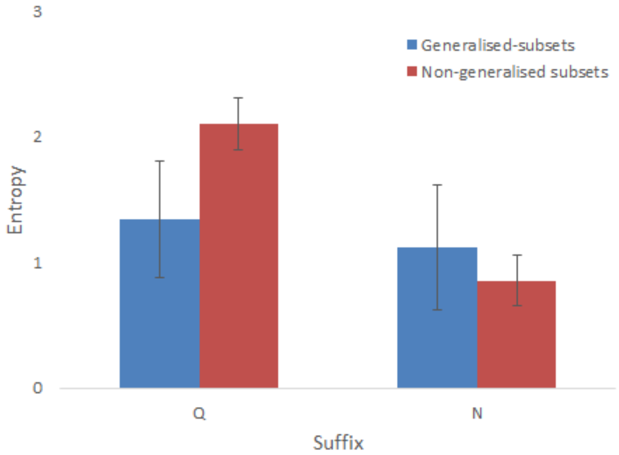


Figure 26: Output entropy for Generalised-subsets and Non-generalised-subsets. Error bars are 95% confidence intervals.

For the quantifier suffixes, the entropy of the Generalised-subsets is lower than that of the Non-generalised-subsets (t(6.624) = 3.885, p = 0.007). The Generalised-subset entropy is also significantly lower than that of the original target language (μ = 1.918, t(5) = -3.204, p = 0.023), though there is no difference between the Non-generalised-subset entropy and that of the target language (μ = 1.918, t(4) = 2.414, p = 0.073).

For the noun suffixes, the difference between the Generalised-subsets and the Non-generalised-subsets is not significant (t(5.491) = -1.354, p = 0.229). Neither the Generalised-subsets (μ = 0.918, t(4) = 1.125, p = 0.324), nor the Non-generalised-subsets (μ = 0.918, t(5) = -0.809, p = 0.455) are significantly different to the target language entropy.

The quantifier results indicate that learners who receive a mix of native and non-native input where the non-native input is a simplified, generalised subset of the native do acquire simpler morphology. This offers support for claims of adult learning simplification of complex morphology, and suggest the problem of linkage could be solved by considering its effects on the input for the next generation.

This analysis is entirely post-hoc, the sample is very small, and there is only a significant result for one of the two word types we have considered, however, it would be wrong to place

[36] As would be expected, there were no examples of participants only receiving complex and simplified Generalised-subset input in the Large-HalfNative condition.

[37] Determining what was and was not a simplified Generalised-subset of the complex input for the verb suffixes proved too subjective.

too much emphasis on it. It does raise an interesting point about the effect of adult learning, however: the *type* of simplification may be important. It may also be that the types of simplified morphological systems produced by the participants in Experiment 5 (and used as the input in this study) are not particularly representative of the what happens in more naturalistic conditions. The original target language was designed so that the stems were particularly easy to acquire, reducing the need to acquire the suffixes for expressivity purposes. This may have affected the participants' approach to learning the suffixes, and this may not be typical of more natural adult language learning.

There is one social factor which has not been included in this experiment which is worth noting. In specifically isolating and focusing on the linguistic input available in different conditions, I have not considered the effect of speaker identity. In natural language learning, the provider of the input is very likely to be important. In our experiment, our participants had no way of knowing if any variability they detected was within or between speakers, and they had no way of knowing if a given input string was provided by a native or non-native speaker. Speaker identity would also make it clearer to a participant that they were learning a static language. One of the participants did report that they had managed to learn the language before "the rules changed".

It is also possible that simplifications arising from adult learning can spread through language use, in the form of simplifications made by natives when talking to non-natives. We consider this in Experiment 7.

3.4 Experiment 7: adult learning and interaction with native speakers

A speaker typically accommodates to their interlocutor when communicating, making both linguistic and extra-linguistic adjustments to facilitate the interaction (Giles et al., 1991). As with Infant (or Child) Directed Speech, this is particularly evident in Foreigner Directed Speech. I suggest that this native-speaker accommodation may provide a crucial linking mechanism by which adult learner simplifications may spread beyond the individual.

In some cases, it may be that a native language speaker may temporarily adopt and produce the simplifications of an adult learner to facilitate their interaction. If they do so, the frequency of these simplifications in the linguistic input of subsequent learners will increase. This may lower the variability of the linguistic input and make the persistence and propagation of such simplifications more likely. It may also be possible that with sufficient exposure to the simplifications of adult learners, native speakers may appropriate them more permanently and use them in their interactions with other native speakers.

Dale and Lupyan (2012) provide some evidence for such effects of interaction with non-native speakers, finding that adult, native English-speakers who had had a greater amount of contact with non-native speakers showed a greater preference for regularised variants of irregular past tense verbs (e.g. "speeded" compared to "sped"). Here, I also aim to experimentally assess the effect of native-speaker interaction with adult learners, by considering the interaction of speakers of a comparatively complex miniature artificial language with speakers of a simplified variant of that language. First, however, I review the characteristics of different types of Directed Speech.

Infant Directed Speech is more widely studied than Foreigner Directed Speech. It is substantially different to normal speech, usually displaying a lower degree of disfluency, shorter utterance lengths, fewer complex sentences, fewer subordinate clauses, higher pitch, lower speech rate, exaggerated intonation, a greater degree of redundancy in the form of replication, restriction of referral to topics outside of the immediate context, and audience design features to actively involve the child more in the discourse (Pine, 1994; Uther et al., 2007). Typically, Infant Directed Speech is simpler than more standard speech, and appears to be cross-linguistically and cross-culturally prevalent without a conscious acquisition stage necessary for the speaker (Uther et al., 2007).

Infant Directed Speech is argued to have three roles: aiding child language acquisition, engaging infant attention, and an emotional-affective role.[38] Language acquisition is facilitated by enhanced "staging" of the process. Vowel hyperarticulation, for example, may aid phoneme disambiguation (Kuhl et al., 1997), or help infants map sounds to meanings (Graf Estes and Hurley, 2013). The degree of the speaker's departure from more standard speech also appears to be sensitive to the child's level of acquisition. The extent of hyperarticulation, for example, appears to reduce with the hearer's age (Xu Rattanasone et al., 2013).

Foreigner Directed Speech is far less extensively researched (Uther et al., 2007), and what research there is focuses heavily on the English language and Western culture. It appears to have only two of the roles of Infant Directed Speech, however. Though not having the emotional-affective function as Infant Directed Speech, it retains the attention-holding and didactic functions. As the emotional-affective role is primarily controlled by pitch (Uther et al., 2007), the most striking difference between Infant Directed Speech and Foreigner Directed Speech is likely to be Foreigner Directed Speech having more standard speech-like pitch control.[39]

Directed Speech selecting those of the three roles which are useful to the hearer can be better appreciated by also considering Pet Directed Speech, which is proposed to typically have the emotional-affective and attention-holding functions, but not the didactic one. Pet Directed Speech, therefore, typically lacks the vowel hyperarticulation of Infant Directed Speech and Foreigner Directed Speech. This appears to be dependent on the speaker's expectations about the linguistic capabilities of the hearer, however. Comparison of Infant Directed Speech, Directed Speech to a parrot, Directed Speech to a dog, and standard speech presents a decreasing amount of vowel hyperarticulation. This is proposed to be due to speakers believing that infants have high linguistic potential, dogs low, and parrots somewhere in the middle (Burnham et al., 2002; Xu et al., 2013).

Wesche (1994) summarises the main features of Foreigner Directed Speech. As with Infant Directed Speech, these features are more exaggerated when the hearer is less proficient. Speech rate will be lower, with exaggerated intonation and stress on topic nouns, and with more frequent and longer pauses. This leads to more careful (hyper)articulation of underlying vowels and consonant clusters, and avoidance of contractions, which makes word boundaries more clearly defined. Utterances are shorter and syntactically simpler. They are grammatically well-formed, except in the case of fragments used to repeat or elicit information for didactic purposes. Canonical word order is frequently used. Optional grammatical forms are retained and the present tense used more often. Adverbial time markers are preferred, conditional constructions avoided and topics placed at the beginning of utterances to increase salience. Given sentence frames are also used in a formulaic way. For example, when giving definitions, there is repeated use of frames such as "This means..." or "It is a kind of..." (Uther et al., 2007).

[38] Though there is some debate as to whether it actually fulfils all of these functions. Benders (2013), for example, concludes that there is no evidence for a didactic role in a study of Dutch Infant Directed Speech.

[39] Though there is some evidence of exaggerated pitch control in Foreigner Directed Speech (Uther et al., 2007).

With respect to vocabulary, more frequent, neutral, and concrete items are preferred, with idioms and slang avoided. Vocabulary is less varied, with a greater use of copulas, and full noun phrases with proper nouns preferred over pronouns. Lexis is elaborated, with information often restated using synonyms or related words (though this may actually end up confusing the hearer). More questions may be used in place of declaratives in discourse, either in situations of one-way information transfer, or in topic initiations. There is also more repetition, comprehension checks, clarification requests, restatements, expansion of hearer utterances, and closed questions (Long, 1981; Wesche, 1994). A greater amount of non-linguistic support is also evident, such as a greater use of gestures, or, where appropriate, drawings and diagrams (Wesche, 1994).

Foreigner Directed Speech also often has negative connotations for native speakers (Ferguson, 1975) and may be employed to underline status differences between native and non-native speakers (Wesche, 1994).[40] As with Infant Directed Speech, it is thought to involve little additional effort on behalf of the speaker.[41]

An experimental investigation into the role of Foreigner Directed Speech on language complexity was carried out by Little (2011). Participants were taught a language which had two strategies for describing a series of images: a morphological and a lexical. The morphological strategy (which included vowel harmony dependency in the suffixes to prevent a lexical analysis) was the more complex of the two. Participants then interacted with a belief that some of their interlocutors had been taught a different dialect of the language, when they had not. The aim was to see if the more simple, lexical strategy was used to a greater extent in the more Exoteric communicative condition, that where communication was with a partner who had been taught a "different" dialect. There was some evidence of this, but only when the first speaker used a lexical strategy to initiate the interaction; otherwise there was no difference between the conditions. This is some indication of how more simple language is used when a speaker believes they are communicating with a speaker of a slightly different language, however.

While Little (2011) considers two strategies within a stable language, the following experiment considers how a language may change as a result of a complex language speaker communicating with a less proficient user of the language. I consider two conditions: a complex speaker communicating with another complex speaker, and a complex speaker communicating with a speaker of a more regular variant of this language. In the first condition, no change in the complexity of the language is predicted as a result of the interaction (in this respect, this is a control condition). In the second, a decrease in the complexity of the language of the first speaker to accommodate to the less-proficient speaker would demonstrate how simplifications of the type proposed to arise from adult learning could spread.

[40] This relates to Thurston's (1994) suggestion that more esoteric communication may be deliberately employed to exclude out-group members. Both non-Foreigner Directed Speech and esoteric communication have also been proposed as being more complex than Foreigner Directed Speech and exoteric communication, respectively.

[41] Ferguson (1975), arguably the instigator of modern research into "Foreigner Talk" (Ferguson, 1971), did speculate that Foreigner Directed Speech was acquired by American schoolchildren from films and books, but ultimately concluded that "the whole foreigner talk register may be seen as a relatively little used resource of the speech community which is available for rapid development into a pidgin when a particular situation of language contact calls for it" (Ferguson, 1975, p. 11).

3.4.1 Materials and methods

The meaning space was comprised of 9 static images: all combinations of 3 animals (a crocodile, duck, and a dog[42]) with one of 3 arrows indicating different movements (straight, bouncing and looping), as illustrated in Figure 27.

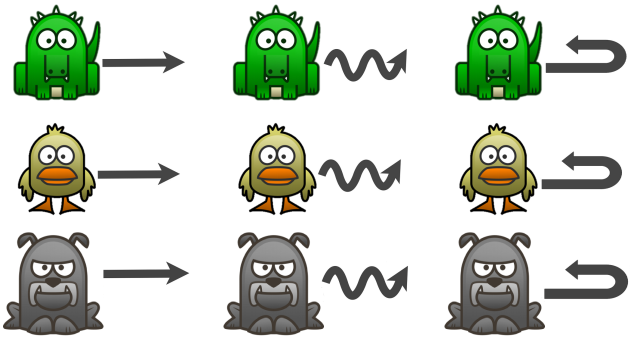


Figure 27: Stimuli set. Made up of every combination of 3 Animals (crocodile, duck, and dog) and 3 Movements (a straight motion, bouncing and looping).

The target language was constructed using 3 "nouns" (each of the form CCVC-*o*) to represent the 3 animals: *snapo*, *kwako*, and *grolo*. As for Experiments 5 and 6, these were intentionally pseudo-English. A given image's label was then made up of one of these nouns and a "verb", separated by a space. The set of verbs was randomly created for each dyad from a set of 5 artificial words (each of the form CVCC): *jing*, *rald*, *nunj*, *ferb*, and *yath*.

3 of these were randomly assigned to the 3 movements, and designated "regular" verbs. The other 2 were "irregular" verbs, and replaced a randomly-selected 2 of the regular verbs, with the stipulation that there would be at most one irregular verb for each animal, and at most one for each movement. An example target language is given in Table 8.

Table 8: Example Target Language. The two irregular verbs are highlighted in red.

Image	Label
crocodile straight	snapo jing
duck straight	kwako rald
dog straight	grolo jing
crocodile bounce	snapo yath
duck bounce	kwako yath
dog bounce	grolo nunj
crocodile loop	snapo ferb
duck loop	kwako ferb
dog loop	grolo ferb

The experiment had two conditions, and the training a participant received depended both on the condition of their dyad, and the role they were allocated within that dyad. In a *Complex-directed* dyad, both participants were trained on the full target language. In a *Simple-directed* dyad, one speaker was again trained on the full language, while the other only received the

[42] The bird, *twito*, of Experiments 5 and 6 was replaced by a dog, as in piloting this study two participants appeared to find the duck and the bird too similar.

7 regular image-label pairings in training. In the Table 8 language, for example, the pairings for “duck straight” and “dog bounce” would be withheld, and so this learner would infer and acquire a regular language where “straight” is expressed as *jing*, “bounce” as *yath*, and “loop” as *ferb*, regardless of the animal in the image. The experiment had 3 stages: a training stage, an interaction stage, and a final testing stage.

Participants were trained to criterion. The participants were first taught the 3 nouns in isolation. As for the training items in Experiments 5 and 6, each image was presented for 1 second in isolation, before the label was also presented as text for a further 5 seconds. The participants then had to retype the label from memory. Blank labels were not accepted, and only the 26 letters, along with commas, spaces, and periods were permitted. No feedback was given as to the accuracy of the retyping.

After this initial pass of the set of nouns, the participants were presented with a training regime which alternated between one of two training contexts. In one, they were either given the pairing and required to retype the label as above. In the other, they were presented with the label alongside the full range of possible stimuli (in this case the 3 animals), and were required to select the correct one using the mouse. Feedback was given as to whether their choice was correct or incorrect, and then the incorrect images were removed so that the correct pairing was made explicit.

After this second training pass, the participant was given a test, in which they were required to label all 3 nouns one after the other. Blank labels were not permitted and no feedback was given. If *both* participants reproduced the nouns with 100% accuracy, the training moved on to teach the noun-verb labels paired with the animal-movement images. If one or both of the participants were inaccurate to any extent, alternating-training phases and testing were repeated until the nouns had been learned. At the end of each test, the quickest member of the dyad had to wait for their partner to finish their test before they could continue.

Training for the noun-verb labels and animal-movement images followed the same regime: first there was an initial retyping pass, followed by alternating-training phases and testing until the participants had reached criterion. For participants trained on the full language of 9 pairings, this was 100% accurate reproduction of the verbs of the target language. For participants trained on only the 7 regular pairings, this was reproduction of the regular verbs. They were not required to label the 2 images which they did not encounter in training.

For both the training of the nouns and the full noun-verb labels, the number of tests was capped at 10. If either participant did not reach criterion by the tenth round, they progressed to the interaction stage anyway but the data from that dyad was excluded from the analysis.

Training was followed by 3 rounds of interaction, in which the participants were instructed to “Try to get as high a score as possible!”. In each round, the participants in a dyad took turns directing and matching labels, with each directing all 9 images in a random order. When asked to direct, the participant was shown an image and required to produce a label. When asked to match, the participant received the director’s label and was required to match it to the correct image which was selected using the mouse. All 9 images were displayed on the matcher’s screen in a random order. Both directer and matcher then received feedback. A green “Success” screen or a red “Failure” screen was displayed, along with a cumulative success score, “Score: x out of y”. No further feedback was given as to which image was intended or selected. There was no indication of a break between rounds, and so the cumulative score displayed to the participants could reach a maximum of 36.

In the final solo testing, the participants were again required to produce labels for each of the 9 images. This allows us to see if any accommodation behaviour exhibited by a participant during the interaction was retained.

The experiment was written in PsychoPy 1.83 (Peirce, 2007).

3.5 Post-experiment questionnaire

Participants were asked to complete a post-experimental questionnaire, comprised of the follow questions:

1. Do you think there were any differences between the language you were taught and what your partner was taught? If so, what were they?
2. Were there any problems communicating with your partner? If so, what were they and how did you solve them?
3. How well did you know the other participant before the experiment?
    - Relationship to you
    - Do you chat to them online (on Facebook, Skype etc.)? If so, how often?
4. Your degree programme (if current or former student)
    - Level (BA, PhD, etc.)
    - Subject

They also completed the Ten-Item Personality Inventory (TIPI) from Gosling et al. (2003).

Degree programme was recorded in case there was any effect of students of linguistics or languages, for example, performing differently to other participants. The TIPI was collected to assess if more “Agreeable” participants were more likely to accommodate in the Simple-directed condition, for example. These variables appeared to have no effect on participant performance, and so they are excluded from the analysis below.

3.5.1 Participants

60 native English speakers (17 male; aged between 18 and 34, mean 19.8) were recruited at the University of Stirling. Each was compensated either 2 Psychology course tokens and £3, or £7. The experiment was run between 4th February and 25th April 2016.[43]

10 dyads involved interaction between participants training on the full language, and so provide data for 20 participants in the Complex-directed control condition. 20 dyads involved interaction between a participant trained on the full language and a participant trained on the restricted set of the regular stimulus-label pairings. These provide data for 20 participants in the Simple-directed conditions, along with data for their 20 restricted-language trained interlocutors.

[43] Data for 6 further participants (3 dyads) was also collected, but it is not included in this analysis as one of the participants failed to meet the training criteria in the maximum 10 rounds of training and testing.

3.5.2 Analysis and results

3.5.2.1 Acquisition of the training data

The training of the nouns in isolation was predictably trivial. 53 of the 60 participants reached criterion in the first test; the remaining 7 took 2 tests. The 40 participants trained on the full language took between 2 and 10 tests to acquire it; average 4.9. The 20 participants trained only on the regulars took between 1 and 9 tests to acquire the restricted language; average 2.1 rounds.

3.5.2.2 Communicative success

Average communicative success across all the interactions was 94%. The average communicative success scores by round and condition of the test items — the irregular image-label pairings which have the potential to be regularised — are shown in Figure 28. On average across all rounds, communicative success for these test items was 98% for the Complex-directed participants, and 74% for the Simple-directed participants. When the Simple-directed participants were matching the labels sent by their partner, communicative accuracy for test items was 97%.

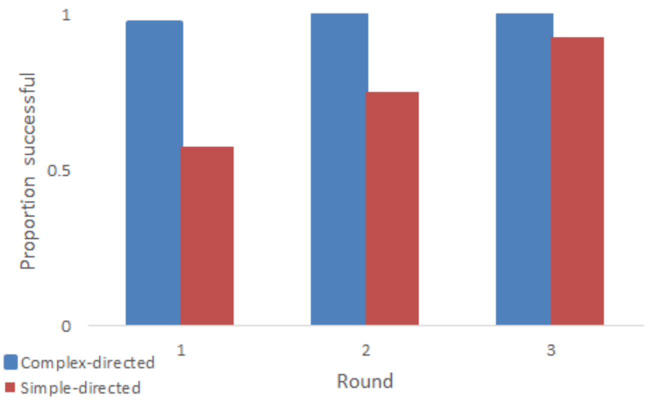


Figure 28: Proportion of successfully communicated test items by round and condition.

A linear mixed effects model with communicative success (1 if the matcher correctly identified the target image; 0 otherwise) as dependent variable included *condition* (Complex-directed or Simple-directed), *round* (which was centred) and their interaction as fixed effects. Complex-directed was taken as the baseline condition. *Participant identity* was investigated as a random effect. This model was significantly better than the equivalent null model ($\chi^2(3)$ = 37.386, p <0.001). There were significant effects of *condition* (β = 3.267, SE = 0.972, p = 0.001), and *round* (β = 1.047, SE = 0.323, p = 0.001). There was no effect of their interaction (β = -0.193, SE = 1.080, p = 0.859).

Unsurprisingly, given the differences in their training, the Simple-directed dyads were less successful at communicating the test items than the Complex-directed dyads. Communicative success improved with round, however.

3.5.2.3 Regularisation in interaction

Over all 3 rounds of interaction, Complex-directed participants on average regularised the irregular verbs 3% of the time. In the language in Table 8, for example, this could be by directing the label “kwako jing” to communicate the “duck straight” image. Simple-directed participants regularised 50% of the time. The proportion of regularised verbs directed by participants who were trained on the full target language including the irregulars by round is shown in Figure 29.

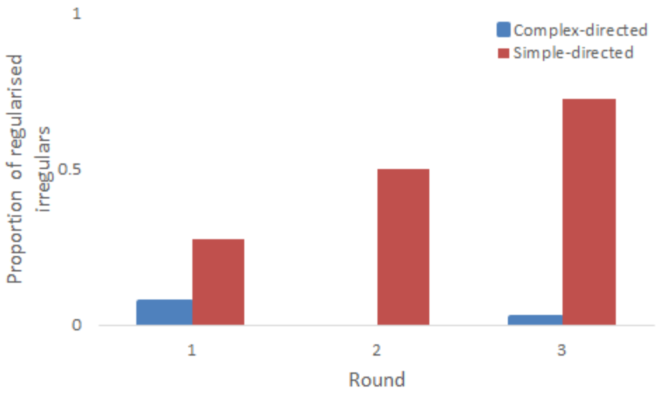


Figure 29: Proportion of regularised irregulars in interaction by round and condition.

Linear mixed effects analysis with regularisation (1 if test item was regularised; 0 otherwise) as dependent variable included *condition* (Complex-directed or Simple-directed), *round* (centred), and their interaction as fixed effects. Complex-directed was taken as the baseline condition. *Participant identity* was investigated as a random effect. This model was significantly better than the equivalent null model ($\chi^2(3) = 53.615$, p <0.001). There were significant effects of *condition* (β = 5.034, SE = 1.176, p <0.001), *round* (β = -1.654, SE = 0.415, p = <0.001), and their interaction (β = 2.599, SE = 0.880, p = 0.003).

Participants in the Simple-directed condition therefore regularised a greater number of the irregular verbs than those in the Complex-directed condition, and there was a greater extent of regularisation in later rounds.

3.5.2.4 Regularisation in individual testing

7 of the 20 Simple-directed participants produced regularised forms instead of at least 1 of the 2 irregular verbs in individual testing, compared to only 1 of the 20 Complex-directed participants. The proportion of regularised irregulars in individual testing by condition is shown in Figure 30. Simple-directed participants regularised irregulars more than Complex-directed participants (t(38) = -2.796, p = 0.008).

3.5.3 Discussion and conclusion

In regularising the irregular forms to successfully communicate with their partners, the Simple-directed participants simplified the target language in interaction. Such simplification was not necessary in the Complex-directed condition, as both participants in the dyad had learned the same full language which included the irregular verbs. In demonstrating that individuals simplify their language to aid communication with interlocutors who have received less linguistic input, this experiment demonstrates how native speaker accommodation to adult learners may lead to language simplification: in increasing the frequency of simplified features, it will increase

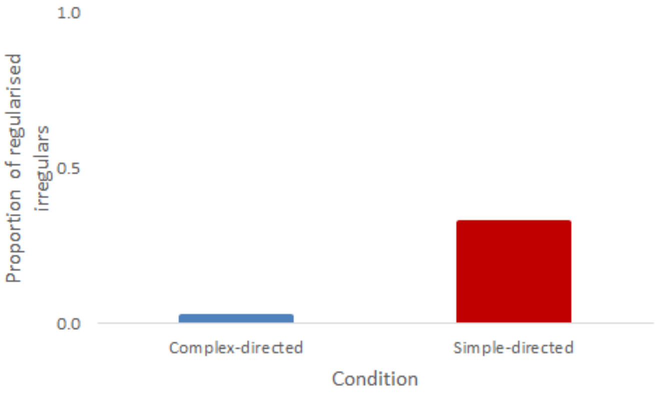


Figure 30: Proportion of regularised irregulars in individual testing by round and condition.

their salience in the linguistic input of other learners, be they children or adults. Unlike the idiosyncractic simplifications of Experiment 6, these simplifications will be more standardised, in that the same simplifications are used by both the accommodating native speaker and the accommodated adult learner. Therefore native speaker accommodation may be a key linking mechanism by which the simplifications of adult learners spread.

This experiment has also provided some evidence of how native speaker accommodation may lead to the spread of adult learner simplifications by horizontal transmission. After simplifying their language in interaction, many of the Simple-directed participants retained those simplifications in individual testing. We may expect such transmission to be minimal in a more naturalistic setting, however. Even if a native speaker retains simplifications for a short period after interaction with an adult learner, we may predict that subsequent interaction with other native speakers eliminates the effect.

It is worth noting two possible objections to these conclusions. The first relates to the restricted input, and the irregular items being inaccessible or less likely to be acquired by non-native speakers. Irregulars are typically high frequency, and so are likely to be present in a learner's input (Cuskley et al., 2015). However, in the acquisition of a much larger language, we may predict that our effects of native speaker accommodation will still influence lower frequency irregulars, redundant features, or in more complex languages than those which have a comparatively low number of particularly salient irregulars. This could be assessed in future research.

Finally, in producing multiple labels for the same stimulus in the Simple-directed condition, it could be argued that this variability actually *increases* the complexity of the language. This increased complexity arising from variability is likely to be short-lived, however, as learners eliminate the unpredictable variation (Hudson Kam and Newport, 2009; Fehér et al., 2014). Again, this could be assessed in future experiments.

3.6 Conclusion

Experiment 5 demonstrated that adult learners, with reduced exposure to the target language, acquire simpler morphology. As non-native speakers are unlikely to reach native-like competence (Selinker, 1972), it therefore seems likely that adult learners acquire and produce simplified language. Even if they do eventually acquire the language as well as a native speaker, there will be a considerable period in which their linguistic output is simplified.

Experiments 6 and 7 consider the problem of linkage (Kirby, 1999), and the mechanism by which the individual-level simplifications of adult learners simplify the language at group level. In Experiment 6, we found that mixing the input from multiple speakers nullified the simplifications introduced in Experiment 5. While the outputs of individual adult learners may be simplified, such simplifications tend to be idiosyncratic, and therefore mixing the output of one or more simplified languages with complex language yields a system which is itself complex and variable. This did not lead to the condition-dependent differences in the languages acquired by our second generation. We did, however, see some evidence for certain types of simplification in the input reducing the complexity of the languages acquired from mixing it with complex input, specifically when the simplified input language is a generalised subset of the complex language. It may be that such systems are more likely to emerge in more naturalistic adult language learning situations, in which acquisition and use are less distinct than we have considered it in this experiment, and so where at least some of the variation between adult learners may be eliminated through interaction. Languages in which the redundancy is not quite as functionless as in the experiment may also reduce the tendency for learners to acquire forms without mappings those forms to meanings. If this is the case, a more discernible simplified system may be apparent alongside the complex. Child learners may eliminate the more complex input as unpredictable variation (Hudson Kam and Newport, 2009).

Experiment 7 focuses on the interaction between native and non-native speakers, by seeing the effect of complex-language speaker accommodation to adult learning-like simplifications. As complex-language speakers regularise more complex language features to facilitate communication, they may increase the frequency and saliency of specific simplifications in the input for following speakers, leading to their propagation. Alternatively, we have seen some evidence for horizontal transmission, with native speakers adopting the simplifications acquired through interaction more permanently. In such cases, this will clearly allow the simplified system to spread.

In conclusion, adult learning is a plausible explanation for why languages spoken by more people have simpler morphology, but native speaker accommodation to non-natives is a key linking mechanism: idiosyncratic simplifications by non-natives alone does not offer a complete explanation.

4 Esoteric communication

4.1 Introduction

As we saw in Chapter 1, the communication of smaller, denser, more isolated, stable societies of intimates will be more esoteric, and this has been proposed to result in more complex language features. The languages will have less transparent form-meaning mappings, a greater number of irregularities, more morphological categories, and greater amounts of syntagmatic redundancy, as well as having more semantically complex lexical items and a greater quantity of "unusual" or "difficult" phonemes and phonotactics (Thurston, 1994; Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011). We have seen proposals for how more exoteric languages may simplify, and so how more esoteric languages may in contrast *maintain* greater levels of complexity, but how might esoteric communication *increase* complexity?

For Wray and Grace (2007), esoteric contexts, in eliminating pressures for simplification, are those which enable languages to return to a more complex, psycholinguistic "default":

> *"The default state is a product of the peculiar facility of the child to acquire language without recourse to full systematicity, and the pressure to minimise processing effort in production and comprehension by dealing with large units where possible and small units only where necessary."*
>
> Wray and Grace (2007, p. 555)

Languages which are solely learned by children are argued to be efficient, non-transparent and non-compositional. They will largely be a set of holistic signals, lacking internal structural composition, which express semantically complex messages. This will be efficient for the speakers, and due to reliably shared context between interlocutors and the absence of interaction with out-group members, they will be no additional comprehension problems for the hearer (Wray and Grace, 2007).

Trudgill (2011, see p. 91-115) offers more detailed explanations of how esoteric-specific language change could lead to increases in complexity. Without the more rapid changes associated with contact and adult learning (Thurston, 1994), languages are more likely to undergo the slower process of becoming more fusional. This process will be aided by reliably-shared communicative contexts in such groups, as there is no pressure for more transparent form-meaning mappings. Such languages are also likely to have a greater number of irregulars due to sound change. Even regular sound change can lead to grammatical irregularities, but in more esoteric contexts more "unusual" sounds changes are more likely, which are even more likely to lead to irregularisation. Random changes may also have greater effects in smaller groups (Nettle, 1999a).

Sound change may also aid the growth of morphological categories. In having a different effect in different (linguistic) contexts (e.g. causing phonological reduction in some contexts but not others), sound change may result in initially meaningless differences emerging within a grammatical category. Reanalysis of these differences may then lead to predictable, categorical distinctions. Without adult learners, who find morphological categories difficult to acquire, these distinctions will then be maintained (Trudgill, 2011).

Esoteric communication may also increase syntagmatic redundancy. Agreement is the most common form of this type of redundancy, and is typically the result of pronoun grammaticalisation. While the processes involved are not entirely understood, it may be that esoteric communication, or communication between native speakers more generally, is aided by such agreement in making meanings clearer to the hearer. This pragmatic function may be detrimental in the cases of exoteric communication, however, again due to the difficulties features such as agreement can cause adult learners (Trudgill, 2011). It is worth noting, however, that esoteric communication having a greater amount of syntagmatic redundancy does seem counterintuitive (and in contrast to Wray and Grace's, 2007, claim that esoteric communication will lead to more efficient language): if there is a greater amount of information shared between speakers, it could be argued that less redundancy could be maintained, as meanings are more likely to be clear from context.

Esoteric communication may also be a means of defining group membership, at the intentional expense of outsider comprehension. As greater transparency aids adult learners, more opaque lexical items and irregularities will hinder them (Thurston, 1994; Wray and Grace, 2007). See Roberts (2010) for a related experimental investigation.

In summary, more esoteric communication is proposed to not only maintain greater levels of complexity in having fewer simplification pressures, but also create the ideal circumstances for increases in complexity. I assess whether esoteric communication has such effects, and whether this is due to group size, group density, or shared information, in four experiments. In Experiment 8, I assess the effects of group size and density on the communicative development of labels for a set of images. Beginning with random, and so highly complex, initial sets of labels, I investigate the extent that they simplify in three conditions. The extent of simplification may prove to be greater in larger groups, and lower in smaller or denser groups. In Experiment 9, I assess the effect of group size on the emergence of linguistic conventions. Unlike in Experiment 8, where I start with a "maximally" complex language and compare degrees of simplification, participants in Experiment 9 communicate in their native language of English, and I compare the extent to which these conventions increase in complexity with increased use. Experiment 10 adds a condition to Experiment 9, directly comparing groups of equal size, but manipulating the amount of information they share. I assess whether the evidence for Exoteric communication, characterised by the members of the group having lower amounts of shared information, producing less complex or more structured sets of conventions than in an Esoteric condition. Finally, Experiment 11 takes the descriptions produced by the participants during Experiments 9 and 10 and uses naive raters to assess their transparency.

4.2 Experiment 8: group structure and language structure

This study compares the evolution of a miniature language in differently structured social groups. I consider three different conditions: a Dyad condition of 2 interacting individuals, an Esoteric Quad condition of 4 interacting individuals, and an Exoteric Quad condition of 4 individuals, but where interaction is restricted so that each only communicates with 2 of the other 3 members of the group. In having a smaller group size, the Dyad condition is more esoteric than each of the Quad conditions. In being more densely structured, the Esoteric Quad condition where all individuals interact is more esoteric than where the interaction is limited to only 2 other members of the group. The design of this experiment and its analysis is based on Kirby et al. (2008) and Kirby et al. (2015).

Kirby et al. (2008) demonstrated how languages evolve to become more learnable and internally structured through repeated transmission to new learners. In the first of two experiments, participants attempted to learn an artificial language which described a structured 3x3x3 meanings space, constructed of 3 colours, 3 shapes, and 3 arrows indicating movement. The first participant in a transmission chain would received 3 rounds of training on a random, and so unstructured, initial language which described a subset of 14 stimuli of this meaning space. After each round of training, participants were given a test in which they were required to label all 27 stimuli. The final testing set of stimulus-label pairings then became the target language for a following participant, who was trained in the same way. Over 10 generations of learning (and 4 chains of participants), the language evolved to be more learnable, as evidence by lower levels of transmission error. As these languages became more learnable, they also lost expressivity, in that single labels were used to describe multiple stimuli.

In the second experiment, Kirby et al. (2008) introduced an expressivity pressure. The 14 items the participants were to be trained on were filtered, and all but one stimulus-label pairing involving a given label were removed. The languages of this experiment become both learnable and expressive. From the initially random sets of labels with no internal structure, the languages evolved to exhibit signs of internal structure, with the components of each label mapping to the components of the meaning space. The initially random languages will therefore have high levels of (objective, Dahl, 2009) complexity compared to the more compositional languages which emerge later (Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011).

Kirby et al. (2015) extend Kirby et al. (2008) to include an expressivity pressure arising from language *use*, and the effect of different social conditions on the emergence of compositional structure. Rather than a single participant learning and being tested on the language at each stage, pairs of participants were trained and then required to interact, taking turns to describe and match stimuli using the labels acquired in training. There were two conditions: an open group condition and a closed group. In the open group, the productions of one of the pair's participants during the interaction was used as training for a new pair. In the closed group, the procedure was the same, except the same pair of participants was used throughout the experiment. Communicative success was higher and transmission error between rounds lower in the closed group compared to the open group. While the languages in the open group became more compositionally structured, those in the closed group did not. The increased pressure to be learnable in the open condition, as 6 pairs of participants in each chain attempt to acquire the language anew, led to an increase in compositional structure and hence decrease in complexity (Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011). In having less stable group membership and the interlocutors having less shared knowledge of past interactions, the communicative context of the open group is more exoteric. The resultant lower levels of complexity would therefore support the proposal that more exoteric communcation results in simplified language, while the closed group illustrates how more esoteric communication can maintain complexity (Wray and Grace, 2007; Trudgill, 2011).

This experiment is an adaptation of the closed group condition of Kirby et al. (2015). I consider different degrees of esotericity in group structure, manipulating group size and group density, and its effect on the compositional structure of the languages which develop through learning and use. I predict that the language of a larger group will become more compositional than that of a smaller group, and that the language of a more densely interconnected group will become less compositional than that of a less dense group. As I focus on the effects of communication here, rather than learning, the participants are only trained at the start of the experiment, rather than after each round of interaction. To allow for any differences between the conditions to be apparent, there are 12 rounds of communication, as opposed to the 6 of Kirby et al. (2015).

4.2.1 Materials and methods

Participants were asked to learn a novel language, and then use it to communicate within a group. The language described a set of 12 images: all combinations of 3 shapes and 4 textures. Each image also had its own unique appendage, so as to allow both holistic and compositional interpretations of the meaning space. These images were taken from Kirby et al. (2015) and are illustrated in Figure 31.

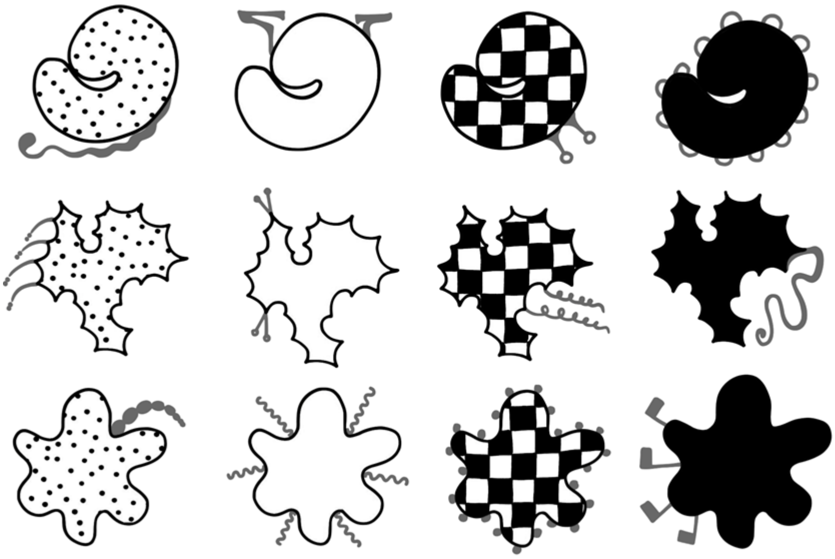


Figure 31: Meaning space for Experiment 8. These 12 images, from Kirby et al. (2015), are all combinations of 3 shapes and 4 textures. Each image is also marked with an idiosyncractic "appendage", so it can be described as either a distinct item of the set or in terms of its shape and texture.

A unique structureless language to describe these images was constructed for each group, following the procedures of Kirby et al. (2008). 12 labels were created by randomly concatenating either 2, 3, or 4 syllables from a random set of unique CV syllables. These labels were then randomly assigned to the the stimuli set.

The experiment involved a training stage followed by 12 rounds of communication, with 3 different experimental conditions. In all conditions, the training stage exposed each participant to 6 randomly sorted passes of the entire set of 12 stimulus-label pairings. Each label was displayed alone for 1 second, followed by 5 seconds displaying the pairing together (as in Kirby et al., 2008).[44] The label was then removed and the participant required to retype it before moving on to the next pairing.

In the Dyadic condition, two participants repeatedly communicated with each other over all 12 rounds. In the Esoteric Quad condition, participants communicated in 2 pairs for each round, swapping partners each round so that each participant communicated with each of the other 3 participants for 4 rounds. In the Exoteric Quad condition, the population structure was

[44] This differs from Kirby et al. (2015) (and Experiments 2, and 4 to 7 in this thesis), however, where the image which was displayed alone for the first second.

less dense, in that not all members of the group directly communicated with one another. The participants again swapped partner for each round, but with one possible pairing withheld for each group member, so that each individual interacted with individuals who had communicative partners they never interacted with themselves. Each participant therefore communicated with 2 other participants for 6 rounds each. The interactive differences in the communication stages are illustrated in Figure 32.

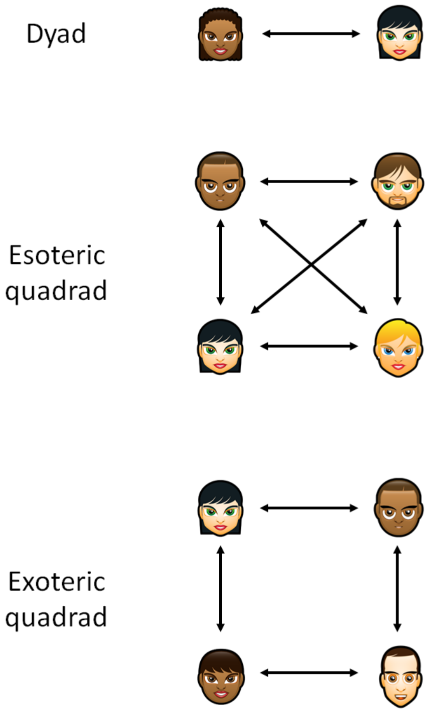


Figure 32: Different social networks of the three conditions. In the Dyad and Esoteric Quad, the network structure is maximally dense, in that all members of the group communicate with one another. In the Exoteric Quad, each participant only directly interacts with two of the other three group members. The network is less densely structured: each individual interacts with others who have a communicative partner who is inaccessible to them.

In a communication round, the entire set of 12 images was placed in a random order. For the first image, one member of each pair was randomly selected as the "director", and the other the "matcher". The director was then shown the the image and required to label it. This label was then shown to the matcher along with the target image and 5 randomly selected foils. The matcher then had to identify the intended image. If they succeeded, the communication was deemed a success; otherwise, a failure. In either case, feedback was provided to both director and matcher, which included the target image, the image selected by the matcher and a cumulative success score for the round so far. Director and matcher would then swap roles for the next image until the entire set of 12 had been communicated. The total set of 12 stimulus-label pairings would then be considered the "language" at that point for analysis.

The experiment was written in Matlab (2010a), and run using Matlab (R2009a), with the Psychtoolbox extensions (Brainard, 1997; Kleiner et al., 2007).

4.2.2 Participants

4 groups were run in each condition, giving a total of 12 groups. 40 native English speakers who were neither current nor former students of linguistics (10 male; aged between 18 and 36, mean 21.7) were recruited at the University of Edinburgh. Each was compensated £10 for approximately 90 minutes of their time. The experiment was run between 25th January and 22nd March 2013.

As part of the application process, participants sent a photo of themselves which was loaded onto the experiment in advance. These photos were then used along with the participant's real names to clearly indicate who was communicating with who in the interactive rounds. Photos and real names were used rather than avatars and other labels so as to make interlocutors more salient and reduce the chance of participants thinking they were being misled as to who they were communicating with.

4.2.3 Analysis and results

Training took between approximately 11 and 19 minutes for the entire group to complete (average approximately 14 minutes). Retyping accuracy for an individual participant ranged between 88% and 100% (binary coded; average 94%). The 12 rounds of interaction took the entire group approximately 56 and 84 minutes to complete (average 65 minutes).

4.2.3.1 Communicative success

Group success scores by round are calculated by dividing the success scores over all communicative partnerships in the group by the maximum possible score (the maximum is 12 in the Dyad condition, but 24 in the Quads). Average communicative success across the conditions rose from 54% in Round 1 to 79% in Round 12. Success scores averaged by condition are illustrated in Figure 33.

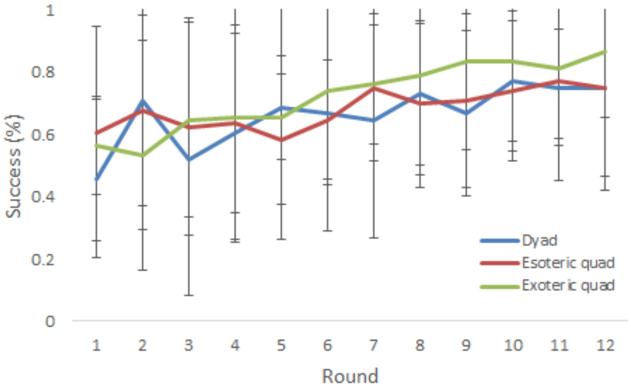


Figure 33: Communicative success scores by round and condition. Error bars are 95% confidence intervals.

Linear mixed effects analysis with communicative success as a dependent variable included *condition* and *round* as (centred) fixed effects, along with *group* as a random intercept effect. This model was significantly better than the equivalent null model ($\chi^2(2) = 66.349$, p <0.001). There was a significant effect of *round* ($\beta = 0.021$, SE = 0.002, t(142) = 9.235, p <0.001), but no effect of *condition* ($\beta = 0.021$, SE = 0.055, t(142) = 0.377, p = 0.707).

4.2.3.2 Structure

The measure of compositional structure follows that of Kirby et al. (2008). For each language, each stimulus-label pairing is systematically compared to each of the other pairings in the language, and two distance measures calculated.

We measure distance between stimuli using the Hamming distance based on their features (shape and texture): this is 1 if the stimuli differ by one of the features (i.e. shape or texture, but not both) and 2 if they differ on both features. The second distance compares the labels, and is calculated using the normalised Levenshtein distance.

For each pair of stimulus-label pairings we therefore have two distance measures. In a structured language, we would expect a greater distance between two stimuli to be reflected by a greater distance between their labels. Therefore we would see a positive correlation between the Hamming distance between the stimuli and the normalised Levenshtein distance between the labels. The Pearson product-moment correlation was calculated for the complete set of pairwise distance measures to measure the level of structure in the language as a whole.

To test the significance of the structural measure for each language, the labels were shuffled and randomly reassigned to the stimuli set 1000 times. For each assignment, the Pearson product-moment correlation was calculated as above, and a population mean and standard variation calculated for the complete set of structural measures. Assuming a normal distribution, z-scores for a given language's structural measure can then then be calculated. At the 95% confidence interval, any z-score greater than 1.96 suggests that a language has a level of structure you would be unlikely to see if it were randomly-constructed. Structure scores for all of the groups are illustrated in Figure 34.[45]

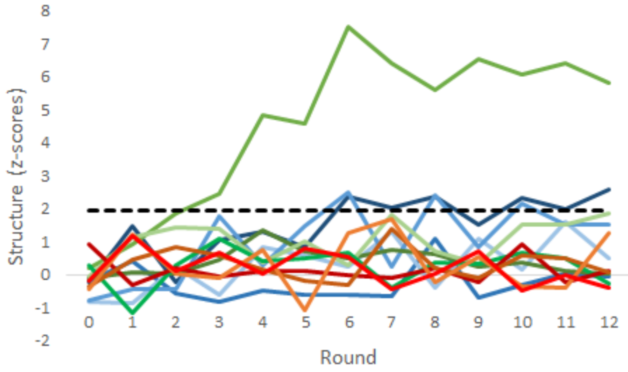


Figure 34: Structural measure for each group (z-scores). Dyads are shown in shades of blue, Esoteric Quads in shades of red/orange, and Exoteric Quads in shades of green. Round 0 represents the randomly generated languages the participants were trained on. The dotted line indicates the critical z-value of 1.96 (95% confidence level). In Round 12, only the languages used by two groups exhibit signs of structure: one Exoteric Quad and one Dyad.

Only two of the groups produced significantly structured languages in Round 12: one in the Exoteric Quad and one in the Dyad condition. Linear mixed effects analysis with z-score as dependent variable again included *condition* and *round* as (centred) fixed effects, along with

[45] For each round, each Quad group produces two languages, as there are two interacting pairs or participants. Z-scores were calculated separately for each language, and then averaged for illustration purposes in Figure 34. This averaging was not done for the statistical analysis.

group as a random intercept effect. The Round 0 seed languages were excluded from this analysis. The model was significantly different to the equivalent null model ($\chi^2(2) = 8.079$, p = 0.018), with AIC indicating a better fit of the data (AIC = 773, compared to 777 for the null model), but BIC suggested the model may be overparameterised (BIC = 790, compared to 787). In any case, there was a significant effect of *round* ($\beta = 0.048$, SE = 0.021, t(238) = 2.349, p = 0.020), but no effect of *condition* ($\beta = 0.713$, SE = 0.458, t(238) = 1.577, p = 0.121). Though there is some indication of the languages increasing in structure over time, there is no evidence of a difference between the conditions. Considering Figure 34 (and the dotplot of the random effect), the increase in structure over the rounds may be overly influenced by the single outlying group in the Exoteric Quad condition.

4.2.4 Discussion and conclusion

There is no evidence of any condition-dependent differences in the compositionality of the languages and therefore this experiment offers no support for the claim that group size or density is a factor in determining the complexity of a language. There is, however, further support here for the findings of Kirby et al. (2015). All 3 of our conditions were closed, in that no new learners attempted to learn the language apart from the initial members of the group. Little or no indication of internal structure then suggests that an increased pressure for learnability may be crucial in the development of such structure; it can not be explained by an expressibility pressure alone. As an increase in compositionality, and so decrease in linguistic complexity (Wray and Grace, 2007; Lupyan and Dale, 2010; Trudgill, 2011), is only observable in Kirby et al.'s (2015) open group condition, this suggests that adult learning may result in language simplification (as investigated in Chapter 3). A lack of adult learning may at least create an environment where complexity can be *maintained*, even if there is no evidence here that it can be increased (Wray and Grace, 2007; Trudgill, 2011).

I cannot, of course, rule out the null result also being a consequence of the experimental design. The communication systems in the groups are not completely stable, even by Round 12. In Figure 33 we can see that the languages have not been sufficient well acquired for full communicative success, even by the end of the experiment. This can also be seen in Figure 35 (discussed below): the language used in one round of communication is not the same as the previous one. With more rounds of interaction, the languages may have stabilised, at which point differences in compositionality between conditions may have been apparent. Given the lack of difference after 12 rounds of interaction, however, and the vast majority of the languages being non-compositional at this point (Figure 34), this is unlikely.

In the course of analysing the data, two other phenomena became evident: evidence of hearer-specific labelling and "lost" labels of the original training language re-emerging in later communication rounds. Though not directly related to this experiment's focus on group size and density influence on compositionality, they are still relevant to sociocultural determination of linguistic features, and I consider these in the next two sections.[46]

[46] I also investigated condition-dependent differences of the languages' average label length, number of unique characters used, and number of unique labels used. Both can be related to Trudgill's (2004a) claims regarding more esoteric communication resulting in smaller phoneme inventories and longer words. See Section 1.4.

Linear mixed effects was carried out for each of these variables, again with *condition* and *round* as (centred) fixed effects and *group* as a random effect. For average label length and number of unique characters, both models were not significantly different from the corresponding null models ($\chi^2(2) \leq 3.857$, p $\geq$ 0.145). In each case, there was also no indication of either *condition* or *round* having a significant effect ($|\beta| \leq 0.413$, SE $\geq$ 0.006, $|\text{t}(154)| \leq 1.36$, p $\geq$ 0.176).

4.2.4.1 Hearer-specific labelling

12 distinct labels are required to communicate 12 distinct meanings unambiguously. 3 of the 4 Dyadic groups did indeed use 12 distinct labels in Round 12, which is also the maximum number of labels which can be used in that condition. The other group used one label to describe two different images, and so had a set of 11 labels. In the Quad conditions, however, the average number of labels used per round was 18.2 and 18.3 at Round 12 (with two communicative pairings in each round, the maximum number of labels would 24 — this would imply no overlap between the labels one pair was using and those of the other pair). Linear mixed effects analysis with average number of labels as dependent variable was carried out on the Esoteric and Exoteric Quad data (excluding the Round 0 training data) using *condition* and *round* as (centred) fixed effects and *group* as a random effect. This model was no better than the null equivalent ($\chi^2(2)$ = 1.832, p = 0.400). The model also indicated no significant effect of *round* (β = -0.048, SE = 0.038, t(94) = -1.246, p = 0.216) or *condition* (β = -0.343, SE = 0.750, t(94) = -0.458, p = 0.648). As the number of labels per round was capped at 12 for the Dyad condition, it is not possible to meaningfully compare it to the two Quad conditions here.

While the use of more than 12 labels is mostly due to instability and inconsistent label use due to error, there is some evidence, both in the data and from participant comments after the experiment, that this is also due to some hear-specific labelling. For example, in one of the Exoteric Quads, one participant consistently uses the label *sefise* when communicating with one partner from Round 6 onwards and *sefi* with the other. Though clearly limited, this could be seen as evidence for how larger groups, in using different labels when communicating the same referents to different group members, could have larger lexicons. This would support other claims linking number of speakers and lexicon size (e.g. Bromham et al., 2015, discussed below).

4.2.4.2 Distance from initial language

Trudgill (2011) claims that linguistic change is likely to be slower in more esoteric populations, while Nettle (1999a) has argued that it may actually be faster (though see Nettle, 2012). I look for evidence supporting one or the other of these positions in this data, by assessing how far each language had changed from the original training data in each round. The expectation was that the distances between the initial language and the languages produced by the participants would increase over the rounds, but that there might be a difference between conditions. As in Kirby et al. (2015), a normalised Levenshtein distance between each label and the training label for the same image was calculated, and then averages taken across all of the labels to give a distance score for a language. The distance between each language and the training data by round is shown in Figure 35.

Linear mixed analysis with distance as dependent variable considered *condition* and *round* as (centred) fixed effects and *group* as a random effect. The model was significantly better than the equivalent null model ($\chi^2(2)$ = 9.59, p = 0.008). There was no significant effect of *condition* (β = 0.027, SE = 0.059, t(142) = 0.452, p = 0.652), but there was a significant effect of *round* (β = -0.004, SE = 0.001, t(142) = -3.100, p = 0.002).

Round therefore has an effect on the distance between a language and the training data, but against expectations, it is a *negative* one. Though noting the small effect size, this is evidence that the languages are reverting to the initial target language. This is particularly striking for one Dyad group, with the distance score falling from 0.18 in Round 8 to 0.01 in Rounds 11 and

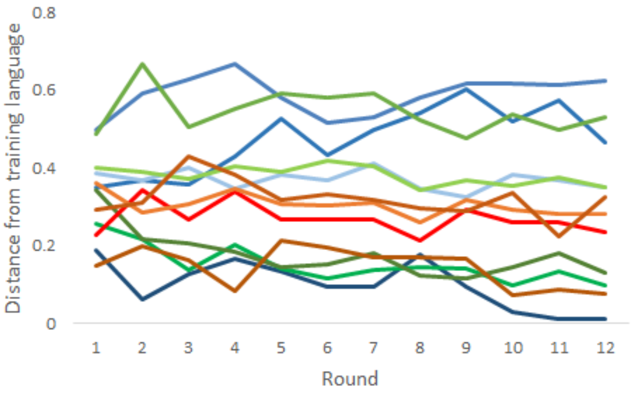


Figure 35: Distance from training language by round. Dyads are shown in shades of blue, Esoteric Quads in shades of red/orange, and Exoteric Quads in shades of green.

12.

The most likely explanation for this effect is that participants, even if they cannot recall a label from the target language, can recognise it as correct if it is produced by another member of the group. Once reminded, they can then use that label when it is their turn to direct. For example, consider one group in the Exoteric Quad condition and a stimulus (the "spotted star" in Figure 31) with the initial language label of *calikete*. In Round 1, both pairs appear to only partly recall the label. In one pairing, the director sends *calili* (which is successfully matched to the target image) and in the other, *keke* (which fails). In Round 2, the director roles swap. In one pair, the correct label of the initial language, *calikete*, is send by both directors. One of these is successful, suggesting that though one of the directors of Round 1 could not recall the label, they could still recognise it and relate it to the correct meaning. In Rounds 3 to 12, the image is identified with the initial label *calikete* 18 times, and a slightly variant of it, *capikete*, twice.

What we see then, is a group's ability to learn a language surpassing that of the individual. This can be related to *swarm intelligence*, where

> "two or more individuals independently collect information that is processed through social interaction and provides a solution to a cognitive problem that is not available to single individuals"
>
> Krause et al. (2010, p. 28)

This has two (admittedly, speculative) implications for us, relating to the open or closed structure of a group, and its size.

If, as in experiments which adopt an iterated learning design (e.g. Kirby et al., 2008), we had included some element of population turnover, the lasting effects of the initial target languages would likely be reduced. Slower rates of change and greater levels of complexity may therefore be expected in more closed, esoteric groups, supporting Trudgill (2004a) and Wray and Grace (2007), rather than Nettle (1999a).

There may also be group-size dependent swarm intelligence effects. Based on the "many wrongs" principle, Simons (2004) puts this in the context of a group of birds attempting to locate an island. The larger the group, the more accurately the goal will be located. As illustrated in Figure 36, a single bird will likely miss the island, but a large group will find it.

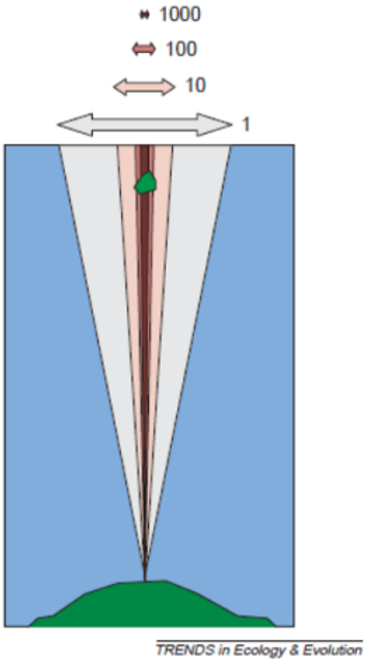


Figure 36: The many wrongs principle. This illustrates the effect of group size on goal achievement. While a single bird (1 member) or a small group (10) is likely to fail to locate the island, membership of larger groups (100 or 1000) increases the chance of success (reproduced from Simons, 2004, p. 454).

We can relate group size and swarm intelligence to the size of a language's lexicon. If larger groups are more likely to retain lexical items even when they are (temporarily) "lost" by individual speakers, this would support previous work linking larger numbers of speakers to languages acquiring words at a faster rate and losing words at a slower rate (Bromham et al., 2015). Maintaining a greater quantity of hearer-specific labels, as discussed above, would also support languages spoken by more people having larger lexicons.

4.3 Experiment 9: group size and the complexity of linguistic conventions

This experiment again investigates the effect of group size and more esoteric communication on the complexity of a miniature language, following previous work in studying the emergence of linguistic conventions as a result of repeating interaction between interlocutors (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986). Unlike in Experiment 8, interlocutors are able to communicate freely in English, and able to ensure communicative success. Feedback is also largely driven by more natural interaction between the members of a group. We consider two conditions: a more esoteric, Dyad condition with a group size of 2, and a more exoteric, Triad condition with a group size of 3. The participants take part in a number of rounds of communication in which they have to label and identify tangrams. Communicative conventions emerge within the group in the form of group-specific labels, and it is these which we assess for structural and semantic complexity.

This experimental design is based on the studies of Clark and Wilkes-Gibbs (1986), which demonstrated the collaborative nature of the establishment of referring expressions, in that both the speaker and the hearer share responsibility for successful communication. In the original study, pairs of participants communicated the same 12 tangrams for 6 rounds. Over the rounds, the lengths of the descriptions decreased, and the number of turn-taking changes between the director and matcher decreased. As an example of the change in the nature of the descriptions, a director described one tangram in Round 1 by saying "All right, the next one looks like a person who's ice skating, except they're sticking two arms out in front." In Round 6, they simply referred to "The ice skater." (Clark and Wilkes-Gibbs, 1986, p. 12).

I extend Clark and Wilkes-Gibbs's (1986) study to investigate the change in complexity of referring expressions in two conditions: Dyadic and Triadic. In each group, participants communicate the identities of 12 tangrams, taken from a larger set which also contains 12 foils. Unlike in the original Clark and Wilkes-Gibbs (1986) study, the tangrams can be considered structured in that each tangram falls into one of four categories. I evaluate the complexity and transparency of the descriptions used by the participants in three ways: length of descriptions, the structure of the set of descriptions relative to the structure of the tangram set, and semantic complexity.

4.3.1 Additional contributions

This experiment was an extension of Experiment 10 (described in Section 4.4), which was carried out in collaboration with Gregory Mills and Kenny Smith. All three of us played a part in the design of that experiment. I adapted the code for this study from that originally written by Greg for Experiment 10.

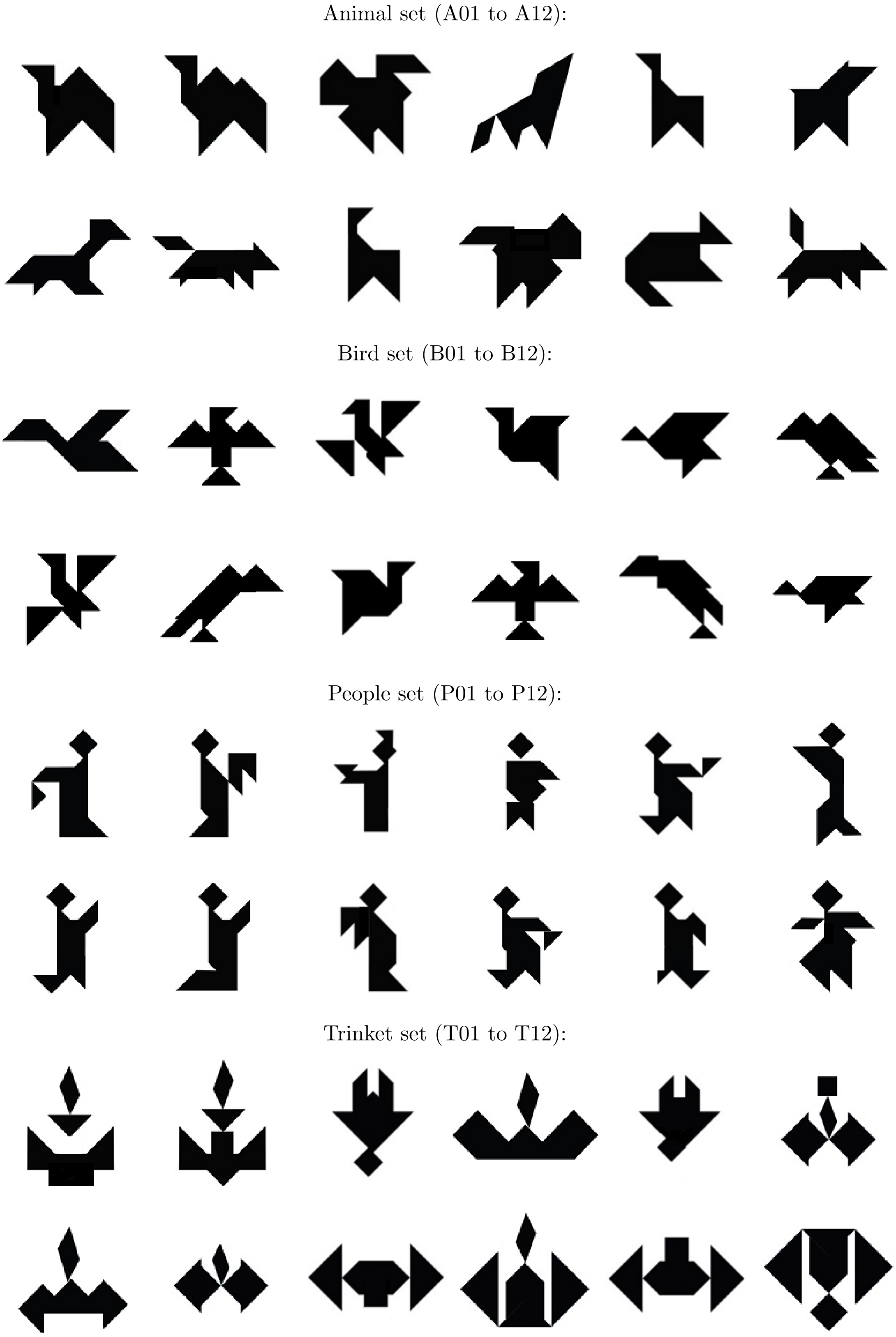
4.3.2 Materials and methods

I constructed a set of 48 tangrams, made up of 4 sets of 12: a set of "animals", "birds", "people", and "trinkets". See Figure 37.

For each Dyad or Triad of participants, 12 tangrams were randomly selected from this larger set as target images, those which would be communicated during the experiment. 12 additional images are selected as foils, which were potential selections for a matcher, but which were never a target for description by a director. The experiment was run using the Dialogue Experimental Toolkit (Mills and Healey, 2016, submitted).[47] This includes an instant-messaging chat-box with which participants could send any message using alphanumeric characters. Message sender was indicated to all participants by the sender's username (selected by the participant), with dialogue history visible. Participants described and matched tangrams over a number of rounds. At the start of each round, each participant was presented with a 6x4 grid displaying the 24 tangrams, presented in a random, participant-specific, order.

In the Triadic condition, the participants were required to identify the same randomly-selected subset of 9 of the 12 communicative images. 3 of these tangrams were randomly assigned to each participant to describe to their partners, and these were marked with a blue border on that participant's screen. Participants were able to select any of the other tangrams in their grid (those not marked with a blue border) using the mouse, and selected tangrams were then indicated by an orange border. These selected images could also be deselected. When all participants had selected exactly 6 tangrams, any participant could use a "Select" button to end

[47] Available at http://cogsci.eecs.qmul.ac.uk/diet/.



Animal set (A01 to A12):

Bird set (B01 to B12):

People set (P01 to P12):

Trinket set (T01 to T12):

Figure 37: Tangram set by type.

the round. Feedback was then given on the directed and selected tangrams. For the directed images, a green border within the blue border indicated that the other two participants had both correctly identified the image; a red border indicated that at least one had incorrectly identified it. For the matched image, a green border indicated that the participant had correctly matched the directed image, while red indicated that they had incorrectly matched the description to the image. Participants were also told the group's score out of 9, where 1 point was awarded for each tangram correctly identified by all 3 participants. The feedback was displayed for 30 seconds at the end of the round. For the final 10 seconds, a "Loading next round..." message was also displayed. An example grid at the feedback stage is illustrated in Figure 38.

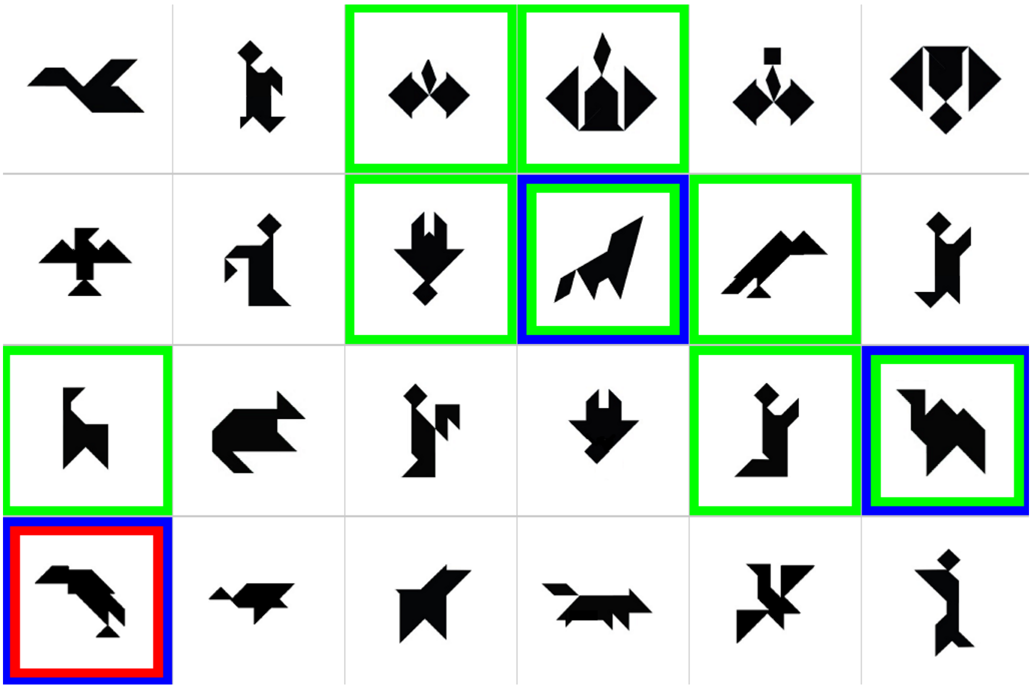


Figure 38: Example feedback screen for participant in one of the triadic conditions. This participant has correctly identified all 6 of the images described to them, as all of the selected tangrams have been marked with a green border. Of the 3 images they described, marked with blue borders, one of them, marked with red, was incorrectly identified by at least one of the other participants.

The rounds in the Dyadic condition followed the same procedure, but with each participant given 4 tangrams highlighted in blue and the aim being to match 8 in total. I aimed to collect a minimum of 6 rounds of data from 10 groups in each condition.

4.3.3 Participants

62 participants (21 male, aged between 18 and 40, mean 21.3) were recruited at the University of Edinburgh. The experiment was run between 6th November 2014 and 11th May 2015. Participants in the Triadic condition were paid £7 for around 60 minutes; in the Dyadic condition £5.50 for around 45 minutes.

4.3.4 Analysis

4.3.4.1 Quantity of data

I collected data from 13 Dyads and 12 Triads. 8 of the Dyads completed 6 rounds of communication and 2 completed 10 in the 60 minutes allocated. The first 6 rounds from these 10 were considered for analysis. The data from 2 Dyads is excluded as they completed less than 3 rounds. Data is also excluded from a Dyad which completed 10 rounds of communication, but did not read the instructions properly and were only guessing what they were supposed to be doing for the first 3 rounds.

10 of the Triads completed at least 6 rounds of communication in the 45-50 minutes allocated, which I consider for analysis. Of these, 5 completed 6 rounds, 1 completed 7 rounds, 1 completed 8 rounds, 1 completed 9 rounds, and 2 completed 10 rounds. Only the first 6 rounds are considered for analysis in each case. Data is excluded from the remaining 2 Triads who completed less than 4 rounds in the time allocated.

4.3.4.2 Line role coding

Each line of text entered during the experiment (a total of 5044 lines) was coded for its communicative "role": either *director*, *matcher question*, *matcher confirmation*, *matcher description*, *turn negotiation*, or *chat*.[48]

Director lines were those used by participants to describe the image they had highlighted in blue. These included responses to matcher questions, included simple confirmations.

Matcher questions included specific direct questions in response to *director* lines, as well as more general requests for further information. They included any lines designed to communicate matcher confusion, need for further details, or elicit some confirmation from the director. For example, "there are two like that", "I think I always describe him as facing left...", or "awesome, I've got it too. It's sort of balancing on a triangle right?".

Matcher confirmations include more implicit confirmations, as well as explicit. For example, "i love that one", or "ah thats a better way to describe it!". They also include less certain confirmations which clearly ended the meaning negotiation process, such as "Hmm I'm just going to guess at one that kind ofseems to fit the description".

Matcher descriptions were only applicable to the Triadic condition, and were used when one matcher, who has already understood a director description, describes the image for the benefit of the second matcher. This includes descriptions by one of the matchers for the benefit of the other, even when in giving the description it is clear that they have not correctly identified the intended director image.

Lines coded as *turn negotiation* included: strategising, e.g. "describe all three at one go", or "Shall we complete one person's set first?"; indications of the participant wanting to assume the director role, e.g. "ok, so mine"; negotiating the end of discussion one image and moving on to the next, e.g. "you two found it?"

Chat marked any line not directly related to describing and matching images or task negotiation, e.g. "alright FU guys i'm having a moment", "YES XXX GODS SAKE ITS BEEN THE SAME ALL ALONG",[49] and "your mom". This included responses to round scores and feedback stages, e.g. "Whoa! Cool!", "imma go go smoke", "fucking candle !!!!", "hashtag

[48] The examples given in this section are taken from either Experiment 9 and Experiment 10. Both were coded in the same way.

[49] To protect participant anonymity, all characters in names or usernames have been replaced with "X".

amazing", and "well done, white boys". It also included any other discussion about the experiment, e.g. "how many rounds are there? I forget", "accidentally hit the wrong one", "click again it will cancel", "dude wtf do we have to do?", and "do you think he reads this?". Lines were also coded as *chat* if they involved extra description or comment about an image after it was clear that all participants felt they had identified the same image, e.g. "another unrealistic body image to be showing young people!" Lines coded like this would always be preceded by at least one *matcher confirmation* line.

For the most part, the coding for each line was relatively trivial. Occasionally, however, the distinction between *turn negotiation* and *chat* was not completely clear, and I had to make a judgement call. Some lines also contain segments which contained elements for two categories. For example, in "yeh i think i see it , the upside down triangle and rhombus are above it?", the first phrase alone would be coded as *matcher confirmation*, while the second would be coded *matcher question*. In such cases, the most appropriate coding in the context of the dialogue was applied. In this example, the line would be coded as *matcher question*, as it elicited further description from the director.

Each *director*, *matcher question*, *matcher confirmation* and *matcher description* line was then marked for the image it referred to. This was done by hand, but with the benefit of knowing which set of images a director had, it was almost always obvious which image was being described. Where a line referred to more than 1 image, it was marked for each image.

4.3.4.3 Isolating descriptions

To analyse the descriptions being used in each round, they first had to be isolated from surrounding linguistic material. To do this, I first separated the *director* lines by image. These lines were then trimmed to only include material which directly described the images. So, for example, "i got the other giraffe" was trimmed to "the other giraffe". End of line commas and fullstops were also removed (any other punctuation was retained). Confirmations or rejections of matcher descriptions, such as "no" or "that's the one!", were removed, and "looks like..." was reduced to "like". Markers of certainty or reference to descriptions in previous rounds were retained. So examples of lines considered (parts of) a description include "the bird i had in round 1 - the one flying up kind of" and "the tissue guy again".

Occasionally, a trimmed description line referred to two images which could not be separated, e.g. "both of the giraffes". In such cases, the description line was considered (part of) the description for each image, but with lexical markers of plurality removed. In this case, for example, the description line would be considered "the giraffes". Similarly, "and then the two bull heads that you guys had" would be taken as "the bull heads that you guys had".

Finally, each character in a participant name or username which was part of a description was replaced with "X", resulting in, for example, "XXXXX's big bird looking to the sky". The trimmed *director* lines for each image were then concatenated to make what we consider the "description" for the purposes of analysis. This approach had the advantage of allowing a high level of consistency in isolating descriptions from the text. One disadvantage is that in the cases where a description was recapped a the end of a round, this is also included in our description for analysis. Another disadvantage is that it makes no consideration of repairs, so both the lines "aztect bird" and "*aztec" are concatenated into a description. Both of these problems may artificially inflate description length, but such instances were relatively rare. The alternative would have been to have had a more subjective approach to deciding when additional or repeated sections of a description were crucial to a matcher's decision as to which image was being referred to.

4.3.5 Results

I analyse the set of descriptions provided by each Dyad and Triad in five ways. First I consider communicative success, measuring how well the descriptions fulfil their communicative function, followed by the lengths of the descriptions, which measures how efficiently they do so. I then have two measures which assess the transparency of the descriptions and the amount of structure in each set of descriptions relative to the tangrams: the similarity of the strings, both within and between the different tangram types they label (animal, bird, person, and trinket), and then the similarity of the semantic concepts those strings represent. Finally, I consider a measure of semantic complexity.

4.3.5.1 Communicative success

Average proportion correct by round and condition are shown in Figure 39. Linear mixed effects analysis with communicative success as dependent variable included *condition*, *round-1* (so that the intercept of the model represents Round 1) and their interaction as fixed effects, with the Dyadic condition as the baseline. *Group* was investigated as a random intercept effect. This model was significantly better than the equivalent null model ($\chi^2(3) = 12.096$, p = 0.007). There was a significant effect of *round* ($\beta = 0.008$, SE = 0.003, t(117) = 2.45, p = 0.016), but no significant effect of *condition* ($\beta =$ -0.003, SE = 0.018, t(117) = -0.16, p = 0.873), or the interaction of *condition* and *round* ($\beta <$0.001, SE = 0.004, t(117) = 0.08, p = 0.936). Therefore though communicative accuracy increased, there was no evidence of any effect of condition.

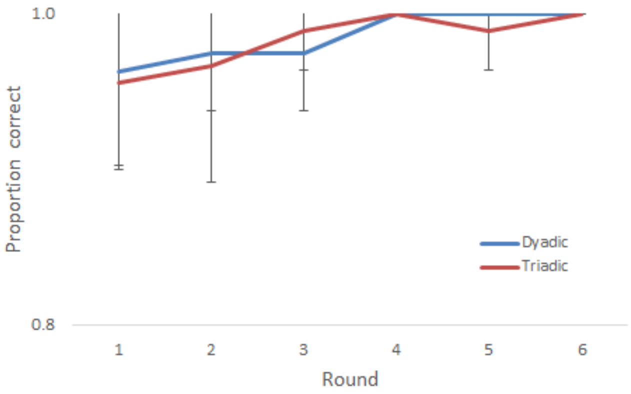


Figure 39: Average scores by condition and round. Error bars are 95% confidence intervals.

4.3.5.2 Description length

Description length captures the efficiency of the conventions (Clark and Wilkes-Gibbs, 1986). There are two approaches to comparing our description lengths by condition. One is to compare sets of descriptions grouped by round, the other is grouped by stimulus occurrence, comparing descriptions for images which have been described the same number of times. I consider both.

Comparison by round ensures that labels produced at the same stage of the experiment are being compared. However, since images can be described multiple times over all rounds but not all these images are described in each round, this may obscure differences which develop as a result of the number of times each image has previously been selected for description. The process of communicating an image marked for direction for the first time in Round 3, for example, displays more of the characteristics of a Round 1 negotiation than that of a Round 3 directed image which was also directed in Rounds 1 and 2. There is also the slight difference in the number of images directed in each round in each condition (9 in the Triadic condition and 8 in the Dyadic), which can be accounted for with occurrence-based description groupings. Figure 40 illustrates the number of image occurrences by condition. Of the 20 groups, 19 have at least 8 images which have been described at least 4 times (1 Dyad described 7 images 4 times). As the figure shows, a much smaller number of images have been described at least 5 times, and so 4 occurrences was deemed a reasonable maximum point to compare the descriptions between conditions.

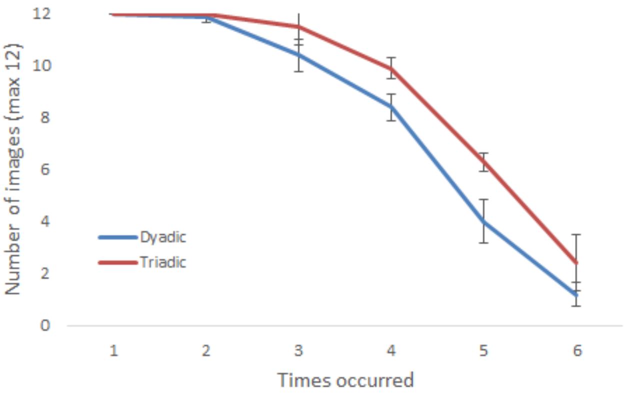


Figure 40: Number of images has occurred at least each number of times by condition. All images have been described at least twice in all but one group. At least 9 images have occurred at least 3 times, and at least 7 images have occurred at least 4 times in all groups. Error bars are 95% confidence intervals.

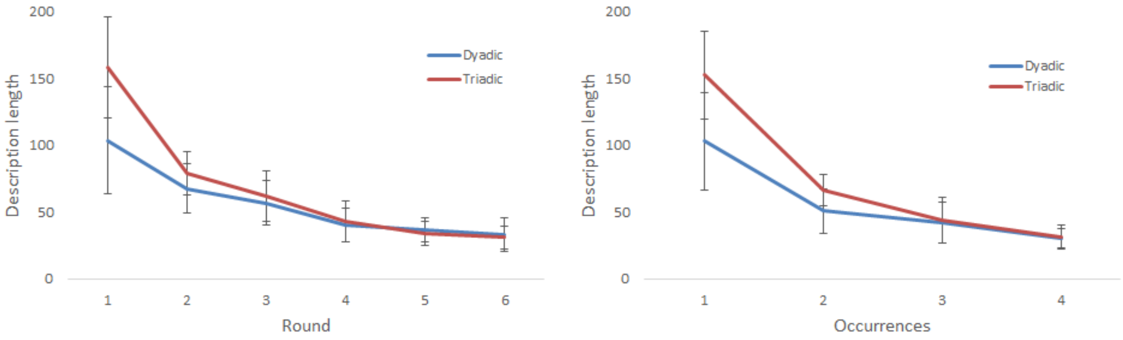


Figure 41: Average length of descriptions by round and occurrence. Error bars are 95% confidence intervals.

The average lengths of the descriptions by round and by occurrence are illustrated in Figure 41. First we consider the descriptions grouped by round. A linear mixed model with description length as dependent variable and with *condition* and *round-1* and their interaction as fixed effects and *group* as a random effect was significant better than its null model ($\chi^2(3)$ = 295.7, p <0.001). There were significant effects of *condition* (β = 35.681, SE = 8.936, t(1018) = 3.993, p <0.001), *round-1* (β = -12.987, SE = 1.483, t(1018) = -8.759, p <0.001), and their interaction (β = -9.632, SE = 2.035, t(1018) = -4.732, p <0.001). While descriptions in Triads are initially longer, they decrease in length more rapidly. Consequently, by Round 6 there is no difference (t(168) = 0.540, p = 0.590). Considering descriptions grouped instead by occurrence shows a very similar pattern of results. A linear mixed model with description length as dependent variable and with *condition* and *occurence-1* and their interaction as fixed effects and *group* as a random effect was significant better than its null model ($\chi^2(3)$ = 341.6, p <0.001). There were significant effects of *condition* (β = 33.959, SE = 8.312, t(1018) = 4.086, p <0.001), *occurrence-1* (β = -17.951, SE = 1.807, t(1018) = -9.934, p <0.001), and their interaction (β = -9.758, SE = 2.407, t(1018) = -4.053, p <0.001). There is no difference between the description lengths of Dyads and Triads in Occurrence 4 (t(181) = -0.297, p = 0.767).

4.3.5.3 String similarity and description transparency

In this section, I take a closer look at the sets of descriptions and consider their relationship to the structure of the tangram sets (the tangrams are structured in that the images are grouped into animals, birds, people and trinkets). Evidence of such a relationship would indicate structure within the set of descriptions and a more transparent labelling of the images, proposed to be more prevalent in a more exoteric communicative context. In this case, therefore, we may predict that the structure of the Triadic sets of descriptions will be greater than that of the Dyadic.

First I consider each set of descriptions by round, and then by occurrence, and again follow Kirby et al. (2008) and the measure of structure described for Experiment 8 (see Section 4.2.3.2 on page 95). I take the normalised Levenshtein distance as the edit distance between pairs of strings, and a binary edit distance between pairs of images: 0 if they are of the same type; 1 otherwise. Monte Carlo simulation uses 10,000 randomised assignments of labels to stimuli, and a z-score greater than 1.96 again indicates that there is a level of structure in the labels which you would be unlikely to see if the descriptions were randomly-assigned to the images (α = 0.05). The structural measures for each condition are illustrated in Figure 42.

As evident from the figure and considering either descriptions grouped by round or by occurrence, the majority of the z-scores for both conditions are less than the critical z-score, suggesting that these sets of descriptions are not typically structured with respect to the tangrams. Some sets (those with z-scores greater than 1.96) do have a degree of structure, however. Table 9 gives an example of a structured and an unstructured set of descriptions. To directly compare the z-scores of the different conditions, a linear mixed model with z-score as dependent variable was constructed with *condition*, *round-1* and their interaction as fixed effects, with *group* as a random effect. This was not significantly different to the null model ($\chi^2(3)$ = 5.861, p = 0.119), indicating no effect of the fixed effects. Considering instead descriptions by occurrence, the equivalent model was significantly different to the null model ($\chi^2(3)$ = 8.651, p = 0.034). The model was a better fit of the data under AIC (257.16 compared to 259.81), but with BIC (266.96 compared to 271.46) suggesting that the model may be overparameterised. In any case, the model indicated that there was no significant effect of *condition* (β = -0.868, SE = 0.445, t(77) = -1.952, p = 0.055) or *occurrence-1* (β = -0.200, SE = 0.154, t(77) = -1.295, p

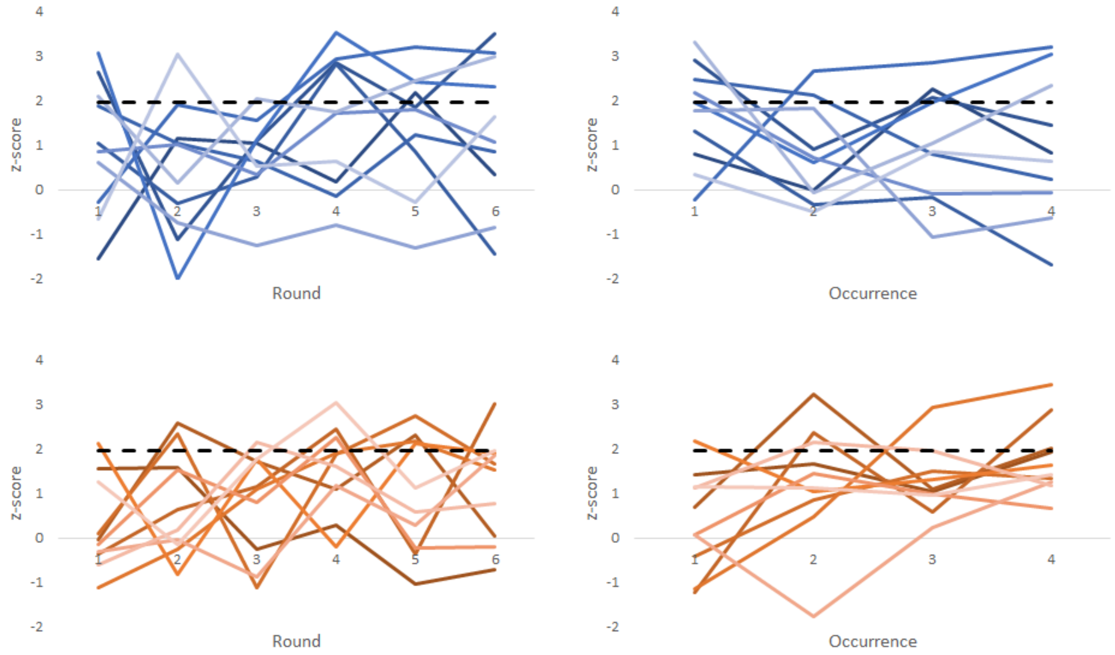


Figure 42: Structure measures based on full descriptions by condition. Sets of descriptions either based on round or occurrence for Dyads (blue) or Triads (brown). Dashed line marks critical z-score.

= 0.199), but that there was a significant effect of the interaction of *condition* and *occurrence-1* (β = 0.615, SE = 0.218, t(77) = 2.823, p = 0.006). The Triadic z-scores do increase with occurrence more than the Dyadic, though given that there is no evidence of the average description set being non-randomly structured, this is not enough to conclude that these sets are more transparent than the Dyadic sets.

Considering the descriptions in their entirety is a little crude, however. An alternative, and possibly more sensitive, measure of similarity between the descriptions is to only consider the "head" word of the description. This head is also a unit we can analyse for semantic relatedness between descriptions, and complexity (see Sections 4.3.5.4 and 4.3.5.5, below).

To isolate the head of a description, I first isolated the grammatical head of the main (i.e. most informative for descriptive purposes) phrase. As it was common and uninformative, the word "one" was ignored. So, for example, in the phrase "animal one", I took the head to be "animal". Where two words could be identified as the head, the first word was taken. For example, in "like an emu or ostrich...", the head was taken to be "emu". Plurals were singularised where the description had originally referred to multiple images (e.g. "men" was coded as "man"), but not where the plurality was part of the description of a single image (e.g. "triangles"). To prepare for the semantic-based analyses which follow below, spelling mistakes were also corrected.

Structure was measured as for the full description strings, and the z-scores are illustrated with the description sets based on round and occurrence in Figure 43. From the figure, we can see that some description sets of heads appear to be non-randomly structured, but not all. Example structured and unstructured sets of heads are shown in Table 10. Comparing

Table 9: Example sets of descriptions. The left set, from Dyad Group 06 Round 1, is structured relative to the set of tangrams (z = 3.074). The right set, from Dyad Group 04 Round 4, is unstructured (z = -0.141).

Image	Set	Description	Image	Set	Description
A04	Animal	a polar bear/lion looking animal	A01	Animal	the camel with one hump
A08	Animal	fox,dog looking animal	A08	Animal	the fox
B06	Bird	an eagle	B04	Bird	the crane
B07	Bird	abird	B06	Bird	the perched bird
P03	Person	man holding bowl	P05	Person	the kneeling person with a triangle pointing up
P05	Person	a man holding a triangle	P06	Person	the jumping dog
P10	Person	on the triangle is pointing down	T03	Trinket	the upside down bird with the diamond head
T03	Trinket	like an upside down person	T09	Trinket	the skull with the recatngle pointing down

Table 10: Example sets of heads. The left set, from Dyad Group 05 Round 4, is structured relative to the set of tangrams (z = 4.434). The right set, from Dyad Group 04 Round 4, is unstructured (z = -0.522).

Image	Set	Description	Image	Set	Description
A10	Animal	plane	A02	Animal	camel
B04	Bird	bird	A04	Animal	wolf
B09	Bird	bird	A08	Animal	raccoon
P02	Person	man	B06	Bird	vulture
P07	Person	man	P05	Person	man
P08	Person	man	P10	Person	guy
T07	Trinket	shape	T01	Trinket	bowl
T08	Trinket	square	T04	Trinket	shape

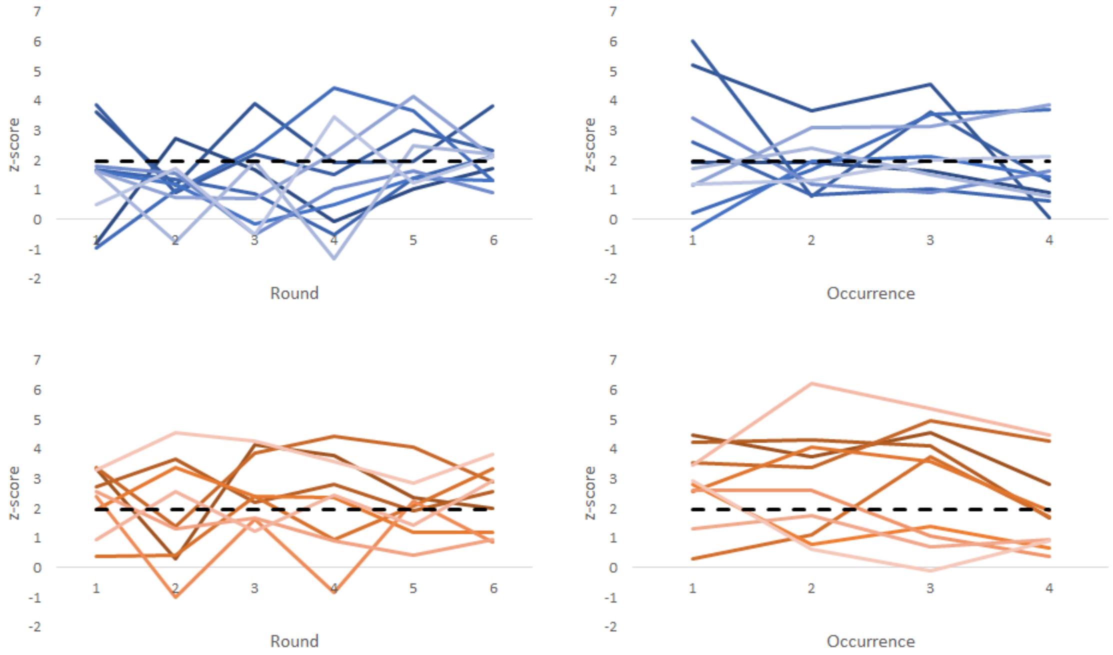


Figure 43: Structure measures based on the description heads by condition. Sets of descriptions either based on round or occurrence for Dyads (blue) or Triads (brown). Dashed line marks critical z-score.

the conditions using linear mixed modelling as before, both the models based on round and on occurrence are worse than the null models, if not significantly so ($\chi^2(3) \leq 5.849$, p ≥ 0.119). There is therefore no evidence that the descriptions in one condition are either more structured or more transparent than the other.

4.3.5.4 Semantic similarity and description transparency

String similarity is not the only way in which a set of descriptions can be assessed for structure. Structure in a set may instead be semantic relatedness amongst the labels for a given category of tangrams, even if the strings are dissimilar. For example, in the unstructured set of heads in Table 10, the Animal set of labels — camel, wolf, and raccoon — are semantically similar in that they are all animals, though they are not string similar. I repeat the above analysis of the sets of heads, but looking instead at semantic distances between pairs of descriptions.

Analysis of the semantic differences between the description heads used WordNet 3.1 (2010). Each unique head (a total of 175 in a list of 1842 heads overall) was checked against its WordNet entry. Where more than one entry existed, the most appropriate was identified. For example, “emu” has 2 definitions, both nouns. The first definition is “electromagnetic unit, emu (any of various systems of units for measuring electricity and magnetism)”; the second is “emu, Dromaius novaehollandiae, Emu novaehollandiae (large Australian flightless bird similar to the ostrich but smaller)”. For the description “like an emu or ostrich sort of facing the leftwith a pointy hump on its back and pointy legsthe body and legs look like an arrow facing upand the top of its head is flat”, I selected the second definition. 5 entries (“batman”, “birdview”, “bob”, “he”, and “toblerone”), the heads for a total of 12 descriptions, had no appropriate WordNet entry and so these were removed from the analysis.

The semantic distance between a pair of heads was calculated using path similarity: the shortest possible hypernym and hyponym path between two WordNet entries. This is scaled so that the maximum similarity between two entries is 1 (i.e. an entry is compared with itself), and the minimum is 0 (i.e. the two entries could not be further apart).[50] Semantic distance was taken as 1 minus path similarity. Where path similarity was undefined, as was the case for pairs of particularly unrelated heads, such as "silhouette" and "blue", semantic distance was taken as 1.

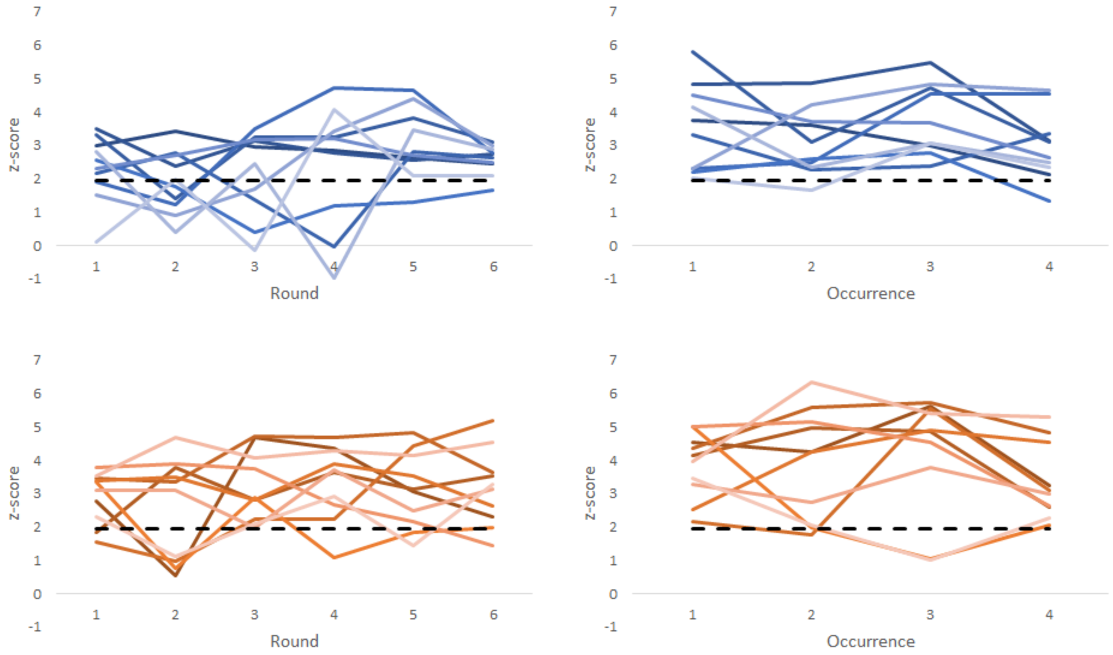


Figure 44: Semantic structure based on the description heads by condition. Sets of descriptions either based on round or occurrence for Dyads (blue) or Triads (brown). Dashed line marks critical z-score.

Semantic structure was calculated using a Mantel test as before. Z-scores by round and occurrence are illustrated in Figure 44. Considering the description sets based on round, a linear mixed model with z-score as dependent variable was constructed with *condition*, *round-1*, and their interaction as fixed effects, and *group* as a random effect. The model was significantly different to the null model ($\chi^2(3) = 9.403$, p = 0.024). Under AIC, the model was a better fit of the data (AIC = 372 compared to 376), while BIC suggests overparameterisation (BIC = 389 compared 384). In any case, none of the fixed effects were significant ($|\beta| \leq 0.783$, SE ≥ 0.403, $|t(177)| \leq 1.944$, p ≥ 0.054). Considering the description sets based on occurrence instead, the equivalent model is no better than the null model ($\chi^2(3) = 2.728$, p = 0.436). We therefore have no evidence for one condition having produced more semantically structured descriptions than the other.

[50] Python implementation available at http://www.nltk.org/howto/wordnet.html.

4.3.5.5 Semantic specificity

Finally, we consider the taxonomic depth of the description heads within the WordNet hierarchy, to assess the claim that more esoteric communication (in this case that of the Dyad condition) will result in greater semantically complexity and more highly-specific lexical items (Wray and Grace, 2007). The average depths for all the heads by round and condition are shown in Figure 45.

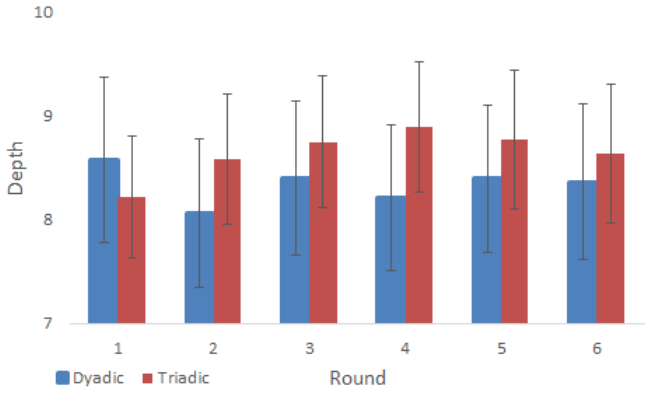


Figure 45: Average WordNet depth of head by condition and round. Error bars are 95% confidence intervals.

A linear mixed model with depth as dependent variable and with *condition*, *round-1*, and their interaction as fixed effects and *group* as a random effect was no better than its null model ($\chi^2(3)$ = 1.830, p = 0.609). Grouping the heads by occurrence rather than round gave the same result ($\chi^2(3)$ = 2.786, p = 0.426). There is therefore no evidence to suggest that the descriptions in one condition are more semantically complex or specific than the other.

4.3.6 Discussion and conclusions

This study has followed previous work in demonstrating that the communication of novel referents becomes more successful and efficient with increased experience (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986; Garrod et al., 2007). Communicative accuracy increases over repeated description of the tangrams, while the length of the descriptions reduced. These effects are also evident using other measures than those discussed above. Round times reduce, as do the number of lines written by the participants. The proportion of *director* lines increases while the proportion of *matcher questions* decrease, suggesting less negotiation is necessary to correctly identify an image from a description. This also replicates Clark and Wilkes-Gibbs (1986), who found that as tangrams were repeatedly communicated, the number of director and matcher turn-taking changes decreased. The number of image de-selections also decreases, suggesting the matchers are more confident in the accuracy of their selections.

Earlier descriptions in the Triadic condition were longer than those in the Dyadic,but with repeated use, they became equally succinct. There was also no evidence of one condition's descriptions being more structured than the other's relative to the meaning space, whether considering similarities between strings or the semantic concepts those strings convey.

The Dyads, in being a smaller group, are thought to have more esoteric communication than the Triads. This has been proposed to result in more complex language. In terms of the descriptions of the tangrams in this study, this complexity would be evidenced by less transparent form-to-meaning mappings, and semantically more complex or more highly-specific lexical items (Wray and Grace, 2007; Trudgill, 2011). If we measure transparency as a greater amount of structure in the set of descriptions relative to the set of tangrams, we therefore have no evidence to support esoteric communication resulting in lower levels of transparency, and this is the case whether we consider string similarity between descriptions, string similarity between description heads, or semantic similarity between description heads. Similarly, there is no evidence of esoteric communication resulting in more semantically complex or more highly-specific lexical items. Therefore there is no evidence of degree of esotericity affecting language complexity here, at least where esotericity is determined by group size.

As with Experiment 8, we cannot of course rule out that our experimental design has failed to capture a genuine effect of group size. The contrast between the conditions may be too slight, for example, and our same measures may capture a difference between the communicating conventions arising from Dyadic or Triadic interaction and those of a much larger group. The inclusion of some element of population turnover may also have exposed a genuine difference between our communicative contrasts (given its effect in, e.g., Kirby et al., 2015).

4.4 Experiment 10: shared knowledge and the complexity of linguistic conventions

I continue the investigation of the effect of esoteric communication communication on language complexity, but investigating the esotericity manipulated by the amount of shared information, rather than group size. I therefore extend the methodology of Experiment 9 to explicitly assess the claim that greater levels of shared knowledge can lead to more complex language. In the previous experiment, all members of the group communicated using 12 of the total set of 48 tangrams, while also sharing the same 12 foils which were never the target of a director's description. To contrast this *Esoteric* triadic condition, exhibiting high levels of shared information, we add an *Exoteric* triadic condition, in which we reduce the amount of shared information by having foils specific to each member of the group. We can then test the hypothesis that lower levels of shared knowledge lead to a set of communicative conventions which display "[m]ore transparency and regularity" and "explicit encoding" (Wray and Grace, 2007, p. 552).

4.4.1 Additional contributions

The experiment was carried out in collaboration with Gregory Mills and Kenny Smith. All three of us were involved in the experimental design. I created the stimuli, piloted the experiment, and created the setup files for running the experiment. Greg wrote the experiment using the Dialogue Experimental Toolkit. I then tested and ran the experiment, and analysed the data.

4.4.2 Materials and methods

As in the Esoteric condition, 12 tangrams were randomly selected for communication, 9 of which were the target for description in any one round. The remaining 36 tangrams were equally and randomly divided between the participants to give each an idiosyncratic set of foils. The methodology for the Exoteric groups is otherwise the same as for the Triadic condition of Experiment 9. Again, I aimed to collect 6 rounds of data for each of 10 groups.

4.4.3 Participants

33 participants (5 male, aged between 18 and 40, mean 22.4) were recruited at the University of Edinburgh. The experiment was run between 30th October 2014 and 1st May 2015. Participants were paid £7 for around 60 minutes.

4.4.4 Analysis

4.4.4.1 Quantity of data

I collected data from 11 groups. 10 completed at least 6 rounds of communication in the 60 minutes allocated, which is considered for analysis. Of these, 5 groups completed 6 rounds, 1 group completed 7, 1 completed 8, 2 completed 9, and 1 completed 10. The final group only completed 3 rounds and was eliminated from the analysis.

All coding and analysis was carried out as for Experiment 9.

4.4.5 Results

4.4.5.1 Communicative success

Average proportion correct by round and condition are shown in Figure 46. A linear mixed effects analysis with communicative success as dependent variable included *condition*, *round-1* and their interaction as fixed effects, with the Esoteric condition as the baseline, and *group* investigated as a random effect. This model was a significantly better fit of the data than the null model ($\chi^2(3)$ = 12.148, p = 0.007). There is a significant effect of *round-1* (β = 0.009, SE = 0.004, t(117) = 2.05, p = 0.043), but no significant effect of *condition* (β = -0.043, SE = 0.027, t(117) = -0.61, p = 0.110) or the interaction of *condition* and *round-1* (β <0.001, SE = 0.006, t(117) = 0.05, p = 0.960). Therefore communicative success increases, but there is no effect of condition.

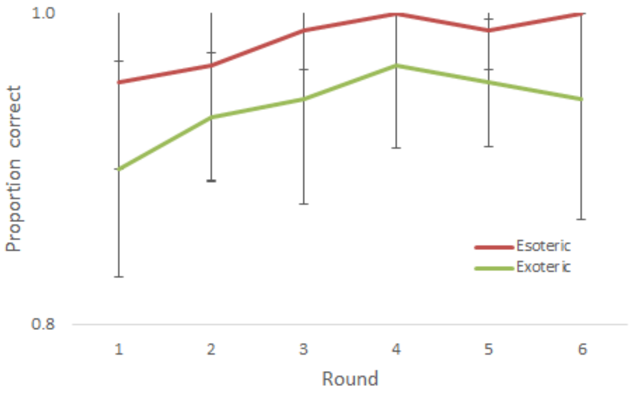


Figure 46: Average scores by condition and round. Error bars are 95% confidence intervals.

4.4.5.2 Description length

As in Experiment 9, we can either compare descriptions by round, or by occurrences for a particular image. All 20 groups have a minimum of 8 images which have been described at least 4 times. The number of images which have occurred at least 5 times across all groups is much lower, however, and so up to 4 occurrences would again seem a reasonable final point for comparing the descriptions across the groups and conditions.

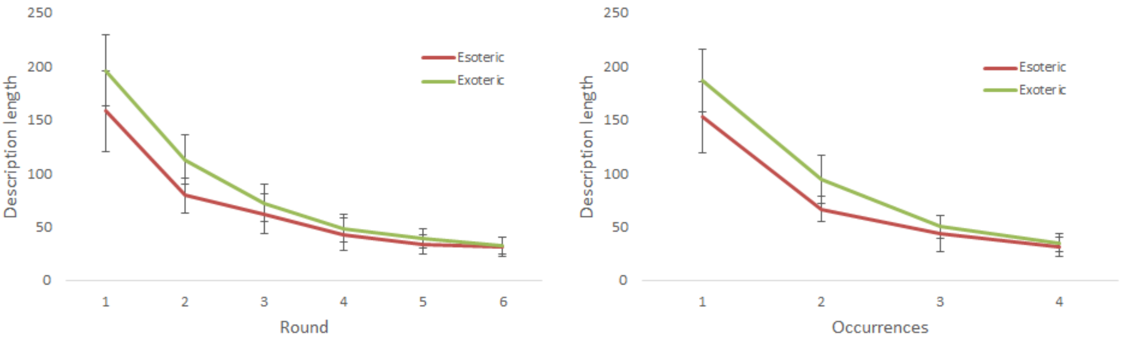


Figure 47: Average length of descriptions by round and occurrence. Error bars are 95% confidence intervals.

Average length of descriptions by round and occurrence is illustrated in Figure 47. First considering the descriptions grouped by round, a linear mixed model was constructed with *condition*, *round-1*, and their interaction as fixed effects, with *group* as a random effect. This fits the data significantly better than the null model ($\chi^2(3)$ = 420.81, p <0.001). There were significant effects of *condition* (β = 35.413, SE = 8.932, t(1079) = 3.965, p <0.001), *round-1* (β = -22.619, SE = 1.685, t(1079) = -13.421, p <0.001), and their interaction (β = -7.878, SE = 1.685, t(1079) = -3.305, p <0.001). While descriptions in the Exoteric groups are initially longer, they decrease in length more rapidly. Consequently, by Round 6 there was no significant difference between the conditions (t(178) = -0.232, p = 0.817). Considering descriptions grouped instead by occurrence shows a very similar pattern of results. The equivalent model again fit the data significantly better than the null model ($\chi^2(3)$ = 389.06, p <0.001), and there were significant effects of *condition* (β = 35.025, SE = 9.541, t(899) = 3.671, p <0.001), *occurrence-1* (β = -39.437, SE = 2.919, t(899) = -13.510, p <0.001), and their interaction (β = -11.456, SE = 4.159, t(899) = -2.754, p <0.001). At Occurrence 4, there was no difference between the conditions (t(190) = -0.755, p = 0.451).

Whichever way the descriptions are grouped, Exoteric descriptions are initially longer than Esoteric, description lengths reduce over the course of the experiment, and that reduction is greater in the Exoteric condition. The difference between the conditions, however, is eliminated as the same images are described multiple times.

4.4.5.3 String similarity and description transparency

Structure measures based on string similarity by round and occurrence are calculated as for Experiment 9. We begin again with the full descriptions, and the z-scores are illustrated in Figure 48. As evident from the figure, there is no evidence that the sets of descriptions are significantly structured on average. A linear mixed effects model with z-score as dependent variable and with *condition*, *round-1*, and their interaction as fixed effects, and *group* as a random effect, is significantly different to the null model ($\chi^2(3)$ = 9.802, p = 0.020). A better fit of the data is indicated under AIC (392.58 compared to 396.39), but overparameterisation implied under BIC (409.31 compared to 404.75). Under the model, there was no significant effect of *condition* or the interaction of *round-1* and *condition* ($|\beta| \leq 0.033$, SE ≥ 0.120, $|t(117)| \leq 0.277$, p $\geq$ 0.782). There was a significant effect of *round-1* (β = 0.173, SE = 0.085, t(117) = 2.036, p <0.001), indicating that the z-scores increase with round. Considering the description sets based on occurrence instead, the equivalent model is no better than the null model ($\chi^2(3)$ = 7.564, p = 0.056). We therefore have no evidence for one condition having more structured description sets relative to the meaning space than the other.

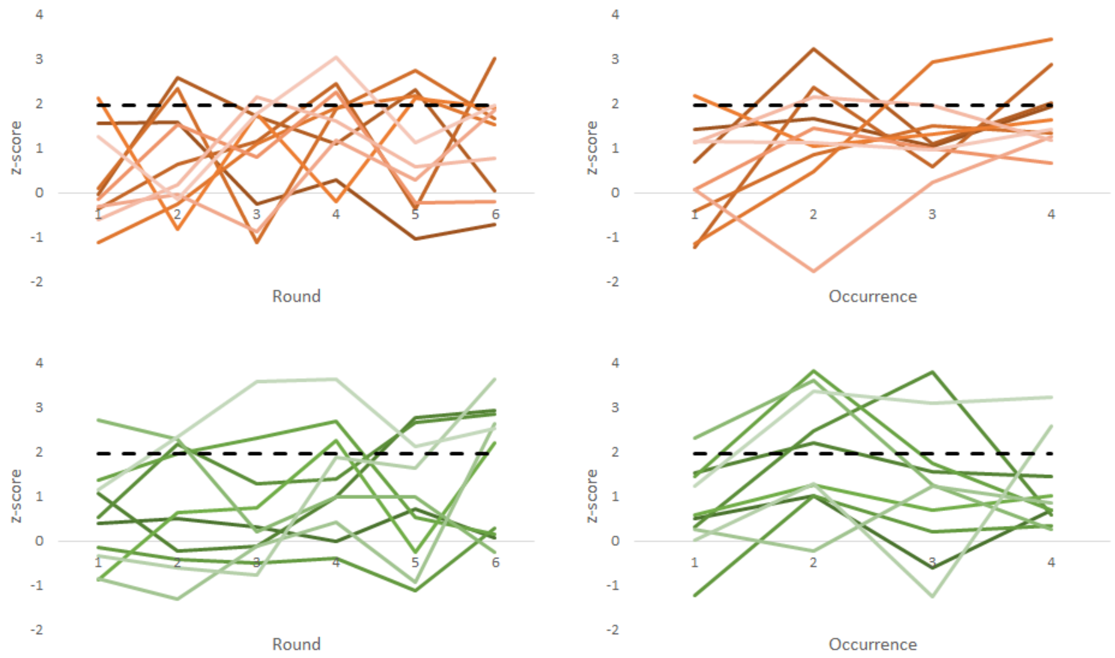


Figure 48: Structure measures based on full descriptions by condition. Sets of descriptions either based on round or occurrence for Esoteric (brown) or Exoteric (green) groups. Dashed line marks critical z-score.

Structure based instead on the heads of the descriptions is illustrated in Figure 49, and as can be seen in the figure, the sets do appear to be non-randomly structured. The linear mixed effects models in this case, considering either descriptions grouped by round or occurrence, are no better than the null models ($\chi^2(3) \leq 2.963$, p $\geq$ 0.397). Therefore there is again no evidence to suggest that the descriptions of one condition are more structured relative to the meaning space than the other.

4.4.5.4 Semantic similarity and description transparency

The structure measures based on semantics are illustrated in Figure 50. The average z-scores by conditions are consistently greater than 1.96 for both the head groupings by round and by occurrence, suggesting the sets of descriptions are significantly structured throughout. Considering the description sets based on round, a linear mixed model with z-score as dependent variable was constructed with *condition*, *round-1*, and their interaction as fixed effects, and *group* as a random effect. The model was significantly different to the null model ($\chi^2(3)$ = 14.17, p = 0.003). Under AIC, the model was a better fit of the data (AIC = 369 compared to 377), while BIC suggests the model may be overparameterised (BIC = 386 compared 385). In any case, none of the fixed effects, *condition* (β = -0.400, SE = 0.436, t(177) = -0.918, p = 0.361),

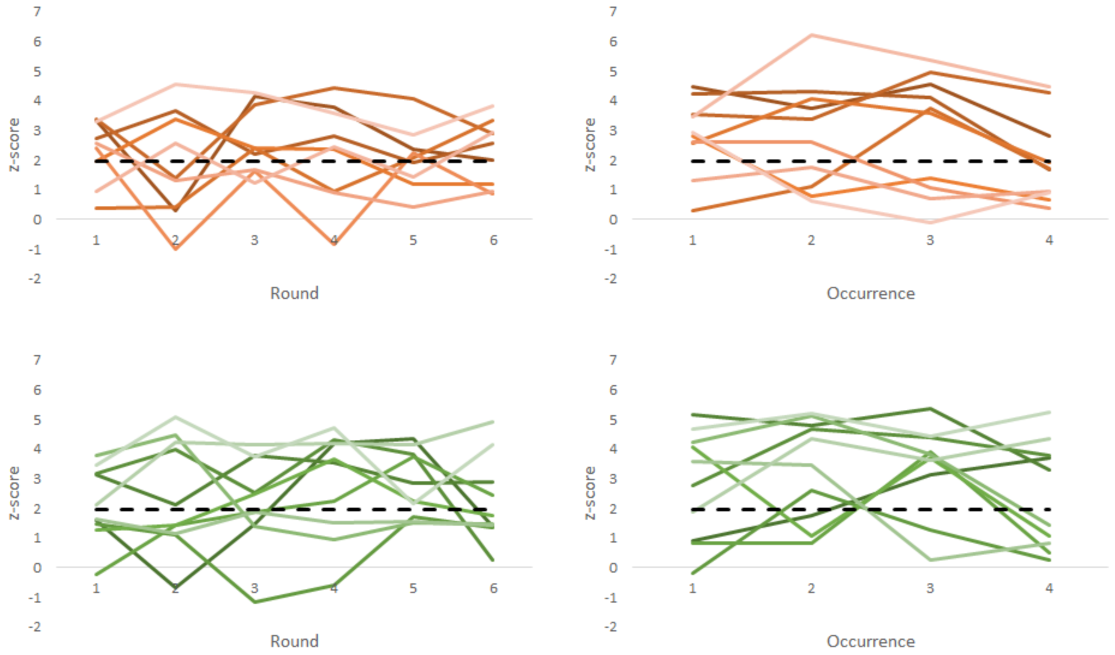


Figure 49: Structure measures based on the description heads by condition. Sets of descriptions either based on round or occurrence for Esoteric (brown) or Exoteric (green) groups. Dashed line marks critical z-score.

round-1 (β = 0.088, SE = 0.074, t(177) = 1.195, p = 0.234), or their interaction (β = 0.182, SE = 0.104, t(177) = 1.749, p = 0.083), were significant. Considering the description sets based on occurrence, the equivalent model was no different to the null model ($\chi^2(3)$ = 0.069, p = 0.875). Therefore there is no suggestion that the Exoteric triads are more semantically structured relative to the meaning space than the Esoteric.

4.4.5.5 Semantic specificity

Finally we consider the semantic complexity or specificity of the heads in each condition, by considering their depth within the WordNet hierarchy. The averages for all the heads by round are shown in Figure 51. A linear mixed model with depth as dependent variable and with *condition*, *round-1*, and their interaction as fixed effects and *group* as a random effect was no better than its null model ($\chi^2(3)$ = 2.830, p = 0.419). Considering *occurrence* rather than *round* gave the same result ($\chi^2(3)$ = 3.901, p = 0.272). Ultimately, there is no evidence of any condition-dependent differences between the labels at Round 6 or Occurrence 4.

4.4.6 Discussion and conclusions

The results of this study largely mirror those of Experiment 9 (see Section 4.3.6), but with the more exoteric condition being the Exoteric Triad over the Esoteric Triad, rather than the (Esoteric) Triad over the Dyad. Again we can see a greater degree of exotericity resulting in longer descriptions in the earlier descriptions, but that difference being eliminating through repeated use, leaving no indication of a lasting influence on the complexity of the set of labels.

The Esoteric Triad, in the group members sharing a greater amount of shared information,

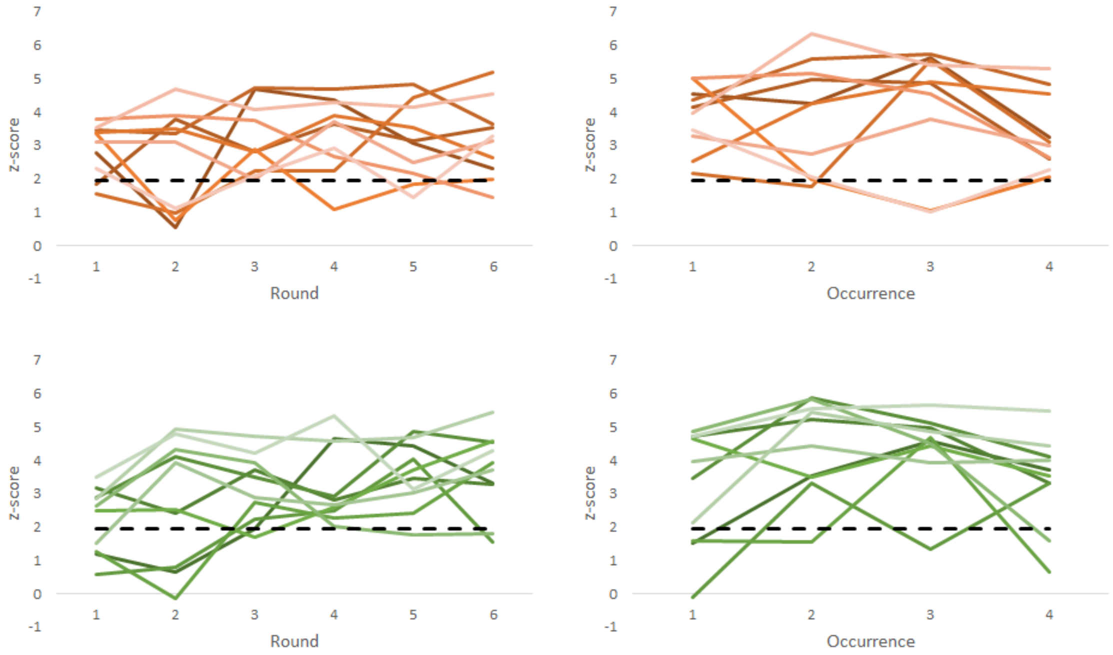


Figure 50: Semantic structure based on the description heads by condition. Sets of descriptions either based on round or occurrence for Esoteric (brown) or Exoteric (green) groups. Dashed line marks critical z-score.

have the esoteric communication, which has been proposed to result in more complex language. As in Experiment 9, this would be evidenced by the sets of descriptions being less transparent, and having semantically more complex or more highly-specific lexical items (Wray and Grace, 2007; Trudgill, 2011). Measuring transparency as a greater amount of structure in the sets of descriptions relative to the sets of tangrams, we therefore have no evidence to support more esoteric communication resulting in lower levels of transparency, and this is the case whether we consider string similarity between descriptions, string similarity between description heads, or semantic similarity between description heads. Similarly there is no evidence of more esoteric communication resulting in more semantically complex or more highly-specific lexical items. Therefore there is no evidence of esotericity, where esotericity is determined by the amount of shared information within a group, affecting language complexity.

4.5 Experiment 11: group size, shared knowledge, and the transparency of linguistic conventions

Finally, I ran an additional experiment in order to test the transparency of the descriptions of Experiments 8 and 9, by seeing how well naive raters could match them to their referents (following, e.g., Fay et al., 2008; Caldwell and Smith, 2012). In removing the shared knowledge established through the grounding of the descriptions, we can more directly assess the claim that more esoteric communication leads to more transparent form-meaning mappings (Wray and Grace, 2007). If we have any evidence in support of the claim, we may expect naive individuals to more accurately match the descriptions produced by Esoteric triads to their intended images compared to the descriptions produced by Dyads. Similarly, they may more accurately

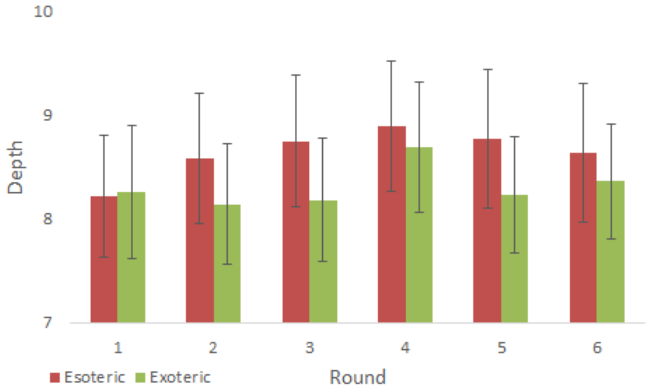


Figure 51: Average WordNet depth of head by condition and round. Error bars are 95% confidence intervals.

match the Exoteric Triad descriptions compared to those of the Esoteric triads.

4.5.1 Additional contributions

This test of the descriptions from the previous two experiments was designed by me, then run on CrowdFlower by Kenny Smith using setup files I provided. I analysed the data.

4.5.2 Materials and methods

I consider the Occurrence 4 descriptions across the 3 conditions of Experiments 8 and 9, with some minor alterations made to the descriptions so as not to unnecessarily confuse the raters. All references to previous labelling of the image or use of the description were removed, including, for example, "AGAIN", "thing XXXX got confused with", "from first round", "we described that one as", and "same". References to participant names or usernames (already marked by a series of "X"s) were removed. Descriptions were de-pluralised where they had been used to refer to multiple images.[51] Where parts of the descriptions were originally written on separate lines, a space was inserted to mark a line break and enhance readability. Finally, 3 labels were excluded in case they caused offence.[52]

The complete set of labels assessed for transparency is given in Appendix E. There are a total of 84 from the Dyad condition, 96 from the Esoteric triad, and 90 from the Exoteric triad: 270 descriptions in total.

I ran 15 trials for each of these 270 descriptions; a total of 4050 testing items. The data was collected on CrowdFlower[53] between 21st July and 7th August 2015. 345 participants were recruited: 330 rated 12 descriptions each, and 15 rated 6 descriptions each. We paid $0.20 for each participant's contribution.

The testing trials were randomly distributed across participants. For a given description, the participant was presented with an array of 24 images, the same seen by a matcher during

[51] In some cases, this highlights a minor problem with this testing procedure. A director in Experiment 9 described images A01 and A02 (the first two images of Figure 37) as "the camels AGAIN". Whereas the matcher's task in the original experiment is likely to have been relatively easy, in this follow up task, the rater is likely to have at most only a 50% chance of identifying the correct image. Examples like this were rare, though.

[52] These referred to A07 in an Esoteric group ("dinosaur with dick out"), P12 in a different Esoteric group ("dancing indian"), and T03 in a Dyad ("the upside down penis person").

[53] http://www.crowdflower.com/.

the experiment. In the Dyadic condition, this is the other person. In the Triadic conditions, one of the two matcher arrays was randomly selected. The arrays were presented in the same order, but what would have been the director images in the experiment were not marked (i.e. this meant that the CrowdFlower participant could select any of 24 images, whereas the participants in Experiments 8 and 9 were not allowed to select the 3 or 4 images they were allocated to direct themselves).

4.5.3 Analysis and results

The average success scores by condition are shown in Figure 52. The Dyadic descriptions had an average accuracy score of 48.9%, the Esoteric 54.2%, and the Exoteric 48.8%. Chance performance was 4.2%. Note that this chance performance is not really comparable to any likelihood of communicative success in Experiments 8 and 9, due to the collaborative nature of the conventionalisation process in these earlier experiments. Here, the recipient of the description had no opportunity to ask for clarification, for example.

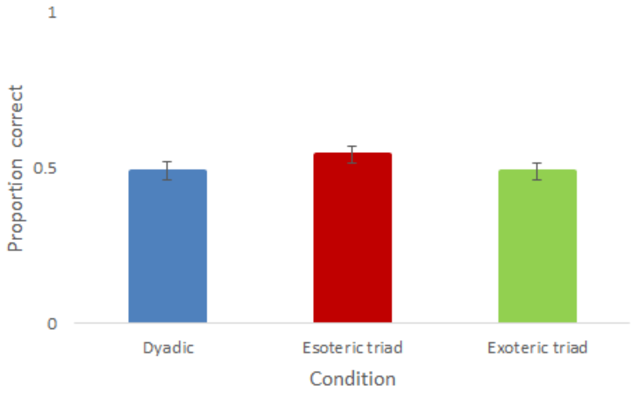


Figure 52: Average proportion correct. Error bars are 95% confidence intervals.

A linear mixed model with logit regression and success score (1 if the target image was correctly identified; 0 otherwise) as dependent variable was constructed with *condition* as a fixed effect and *rater identity* and *intended image* as random intercept effects. The Dyadic condition was set as the baseline. The model was significantly different to the corresponding null equivalent ($\chi^2(1)$ = 8.219, p = 0.017). AIC indicated a better fit of the data under the model (5219 compared to 5223), though BIC implied that the model may be overparameterised (5250 compared to 5242). This is likely an effect of much of the variance in the data being explained by the random effects (0.661 for *rater identity* and 0.568 for *image*). Inspection of the dotplot of random effects did not appear to indicate any outliers, either among the *raters* or the *intended images*, so there is no suggestion of any condition-dependent difference being driven by the more extreme performance of a few individuals, for example.

The model indicated a significant effect of the Esoteric condition relative to the Dyadic (β = 0.269, SE = 0.094, p = 0.004), but not to the Exoteric (β = 0.177, SE = 0.098, p = 0.071). This suggests Esoteric images were 1.31 times as likely to be correctly identified as the Dyadic images. A Tukey multiple comparisons of means then finds no difference between the Esoteric and Exoteric conditions (difference = -0.092, SE = 0.095, p = 0.601). This also confirms that the descriptions from in the Esoteric triad condition were significantly more accurately identified than those of the Dyadic (difference = 0.269, SE = 0.094, p = 0.012).

There is evidence here that the Esoteric triad descriptions of Experiment 9 are more transparent than the Dyadic, but that there is no difference between them and the Exoteric triad descriptions of Experiment 10. Increasing the exotericity of a group by increasing the group size would then appear to have the effect of language simplification, in the form of increasing form-meaning transparency, while an increase in exotericity due to interlocutors having less shared knowledge has no such effect. This is supported by there being no evidence of a difference in transparency between our Esoteric and Exoteric triad descriptions. The lack of a difference between the Dyad and Exoteric triads would appear to go against this hypothesis, however, though note that the model does imply an approaching significant difference (p = 0.071). It is also worth noting that taking the raw average scores by condition, as illustrated in Figure 52, gives a misleading interpretation of the results compared to the model, which also takes into account the large amounts of variance attributable to naive rater identity and the particular image which the description referred to. Both the model and the averages by condition suggest that the Esoteric triad descriptions are more transparent, comparing just the average scores gives the false impression that the Exoteric triad descriptions are less transparent than the Esoteric and as transparent as those of the Dyads.

This study ultimately finds evidence to support more esoteric communication resulting in less transparent form-to-meaning mappings (Wray and Grace, 2007). This effect is only apparent when esotericity is reduced by an increasing group size, not when it is reduced by a reduction in shared information between group members. The increased transparency, however, has only be detectably by asking naive individuals to match the descriptions to their intended referent; as seen in Experiment 9, there is no evidence of the increase in transparency being a result of there being higher degrees of structure relative to the meaning space in the sets of descriptions produced by larger groups.

4.6 Conclusion

In the first three experiments of this chapter, I have illustrated how repeated interaction increases communicative success. In Experiments 8 and 9, we also see increases in communicative efficiency in 3 different conditions, with increased group size and decreased shared information only having an effect on efficiency in the early stages of interaction.

We have very little evidence between our experimental conditions of any effects of esoteric communication on linguistic complexity. Increases in exotericity in the form of increased group size, decreased group density, or reductions in the amount of shared knowledge have no discernible effect on any lexical structure relative to the structure of the meaning space. Group size and shared information manipulations also appear to have no effect on the specificity or semantic complexity of individual lexical items. Considering these results alongside those of Kirby et al. (2015), we may conclude that while there is no evidence that esoteric communication can increase linguistic complexity, the reduced pressure to be learnable (by adults) demonstrates how complexity may at least be maintained (as discussed in the Introduction to Chapter 3 on page 52).

In Experiment 11, by asking naive raters to match the descriptions from Experiments 8 and 9 to their intended images, we find some indication that an increase in group size may result in an increase in transparency, as claimed by Wray and Grace (2007) and Trudgill (2011).

This difference does appear to be specifically dependent on group size, rather than the group members having different background knowledge in the form of different distractor tangrams, suggesting that if shared knowledge is an important factor in determining linguistic communication as proposed (Wray and Grace, 2007; Trudgill, 2011), it may be that it is the shared knowledge *of previous interactions* that is influential, rather than sharing knowledge of sets of potential referents. I suggest that greater amounts of grounding between group members may be the mechanism which explains why esoteric communication leads to higher levels of linguistic complexity.

Fay et al.'s (2008) non-linguistic study which compares conventionalisation of signs in a graphical communication task provides some support for this, finding that signs emerging from a larger group condition prove more transparent than those of dyads. In their study, the signs were equally iconic in the earlier interactions in both types of group, and the transparency effect is due to greater levels of this iconicity being retained over the conventionalisation process in the larger groups. Increased amounts of interaction between the same two individuals in the dyads reduces the pressure for such transparency to be maintained. In our study, we get a similar result, but this may instead be due to there being greater levels of transparency in the larger group's descriptions even in the earlier interactions. The earlier Triadic descriptions are longer, and as they had to be understood by more individuals, they are likely to be easier for naive raters to match to their referents compared to the early descriptions of the Dyads. This could be tested using a new set of naive raters.

Finally, there is also an alternative way of considering esoteric and exoteric communication in these experiments. Rather than focusing on the contrasts between the conditions of our experimental designs, the contrast between the earlier interactions between group members and the later could be compared. In the earlier rounds of all three of the experiments, the participants are in an exoteric communicative context. They have little explicit shared information relevant to the communicative context, are in a contact situation with adults acquiring a novel communication system, and are for the most part, literally "talking to strangers" (Wray and Grace, 2007). By the end of the experiments, they are more like a society of intimates, having increased the amounts of shared information and grounded communicative conventions, and in some cases adapting their utterances to suit specific hearers. Under this contrast of esoteric and exoteric communication in our experiments, we would have evidence of esotericity influences on language features. There is evidence of longer descriptions in Round 1 compared to Round 6 in Experiments 8 and 9. Considering only the Round 1 contrasts, we see that increasing exotericity, either by increasing group size (Experiment 9) or decreasing the amount of shared background knowledge (Experiment 10), further increases the description lengths. This would suggest that the descriptions are more explicit in their encoding of semantic information, claimed to be more representative of exoteric communication and more accessible to non-native speakers (Wray and Grace, 2007). A possible extension to Experiment 11 would be to ask naive raters to identify images comparing Occurrence 1 descriptions across our 3 conditions with Occurrence 4. In doing so, the hypothesis that the earlier descriptions would be more easily identified, and so are more transparent, could be tested.

A key issue with this view of esoteric and exoteric in these experiments is that we would be equating exoteric communication with situations when group members attempt to communicate the identity of novel referents using unconventionalised and ungrounded labels, and esoteric with the communication of familiar referents with at least partly conventionalised labels. Most natural languages would therefore be considered for the most part “esoteric”. As we are interested in the more long-term differences between languages as a result of sociocultural factors such as group size and amount of shared information, interpreting our experimental findings in this way is ultimately inappropriate.

5 Implications for sociocultural determination of linguistic complexity

5.1 Introduction

This thesis has investigated two proposals: languages spoken by larger groups of people are morphologically simpler, and the social situations in which more esoteric communication is employed results in more complex languages. In assessing the first, I have considered two candidate mechanisms by which number of speakers could have such an effect: speaker input variability and adult learning. In the second, I have investigated the effects of esoteric communication by manipulating group size, group density, and amount of shared knowledge.

In this final chapter, I reconsider the literature discussed in Chapter 1 and how sociocultural factors influence linguistic complexity in the light of these experiments. In Section 5.2, I review the direct implications of Chapters 2 to 4 for the proposed effects of social group structure on linguistic complexity, before discussing alternative explanations in Section 5.3. In Section 5.4, I then broaden the discussion to consider the relationship between group size and both other cultural traditions in humans and the communication systems of other species. Finally, I offer some final thoughts and directions for future research in Section 5.5.

5.2 Implications for the field

This thesis attempts to contribute to our understanding of how sociocultural factors may influence language, focusing on how linguistic complexity may be determined by the type of group which speaks the language. In doing so, I have attempted to add to work which may explain the great variation observable amongst languages (Evans and Levinson, 2009), "a core goal of linguistics" (Futrell et al., 2015, p. 1). In focusing on determinants of the structural properties of language, research in this area may be able to shed light on why languages exhibit different degrees of grammatical complexity (Dale and Lupyan, 2012), what the upper and lower bounds of this complexity may be (Gil, 2009; Nichols, 2009), and how individual-level learning interacts with sociocultural features of a speech community to result in group-level language features (Kirby, 1999; Smith and Kirby, 2008; Smith, 2012). It may also aid our understanding of typological and psycholinguistic constraints on language (Wray and Grace, 2007; Trudgill, 2011; Dale and Lupyan, 2012). If the esoteric groups we have been considering are also representative of the sociocultural environments in which language structure first emerged, we may also be able to speculate on the characteristics and development of early language in our species (Wray and Grace, 2007).

Chapters 2 and 3 focused on the observation, evidenced most convincingly by Lupyan and Dale (2010), that languages spoken by more people are less complex. I have assessed two of the mechanisms proposed by Nettle (2012) by which population size could have such an effect. Speaker input variability, as I argue in Chapter 2 and Atkinson et al. (2015), does not appear to explain why languages with a greater number of speakers are more simple. Not only do the results show no indication of input variability having any influence on the acquisition of complex morphology, but the presumption that a learner will receive more variable input in a larger social group, crucially relating to the input relevant to language acquisition, remains to be proven. While the learner may well have a larger social network in larger populations, this does not take into account the relative importance of the individuals within that network. Family size and role of particular caregivers, for example, may be more important factors than network size (Barton and Tomasello, 1994).

A role of adult learning, therefore, would appear to be a more likely explanation (Nettle, 2012). Adult language learning simplification at the level of the individual does appear to be relatively trivial to demonstrate, but following Kirby (1999), I argue that a secondary mechanism which links this individual simplification to the language at the level of the group needs to be demonstrated. Idiosyncractic adult learning simplifications forming part of the input for subsequent generations alone do not generally provide a complete explanation, and native-speaker accommodation may be a key linking mechanism. However, as seen in Experiment 5, simplifications which are specifically simplified subsets of an adult learner's target language may survive cross-generational transmission. More research is necessary to confirm this effect, and determine which of the two types of morphological system acquired by adults is typical in more naturalistic learning contexts. Whether or not such simplified variants are cross-linguistically typical of low-exposure adult language learning requires further investigation, however.

Instead, there may be auxiliary or alternative mechanisms which propagate the simplification effects of adult learning, and these may be sociocultural. In Experiment 6, complex-language speakers simplified their language by regularising irregular verbs to facilitate communication with speakers of a simplified variant of their language. If the complex-language speakers of this experiment are representative of native speakers, and the simple-language speakers representative of adult learners, then such accommodation would increase the frequency of simplified language in the input for subsequent learners. These simplifications could then propagate. Though more limited, there is also some indication of the complex-language speakers appropriating simplified language after interacting with the simple-language speakers. Though more evidence would be needed to demonstrate the extent to which native speakers actually acquire simpler language as a result of interaction with adult learners, this does illustrate how simplifications could spread via horizontal transmission.

Previous research into Foreigner Directed Speech, however, suggests that though native speakers may generally accommodate to less proficient speakers by using simpler language, they still produce grammatical utterances (Wesche, 1994): they do not simplify their language to the extent that they produce utterances which would be considered ill-formed by other native speakers. Given the limited cross-linguistic and cross-cultural research on foreigner directed speech (Uther et al., 2007), however, it may be that native speakers simplifying their language may be more common if the focus of the research is moved further from the study of more modern, European languages (Wray and Grace, 2007; McWhorter, 2009). Either way, the contexts in which native and non-native speakers interact, or the extent of the integration of the non-native speakers into the network structure of speakers as a whole, for example, may be crucial.

In Chapter 4, the focus of my investigation shifted to consider how languages could *increase* in complexity. Four experiments considered how more esoteric group structures or communicative contexts could result in more complex language. The first, Experiment 8, rather than seeing any differences in complexity arising from differences in the experimental conditions, gave additional evidence for the claim that adult learning can cause language simplification. This study can be viewed alongside other work which has seen simpler, more internally-structured language emerge under an increased pressures to be learnable (Kirby et al., 2015). It illustrates how limited interaction with strangers, and so less pressure to be learned and used by adult learners, can preserve complexity. There is no suggestion, however, of smaller or denser social groups maintaining relatively higher degrees of complexity.

Experiments 9 and 10 then largely indicate that manipulations of degree of esotericity, either by social group size or amount of shared knowledge, only produce very short-term differences in emergent linguistic conventions. In Experiment 11, in which naive raters were asked to identify the intended images from descriptions, however, there was some suggestion that a larger group may produce more transparent linguistic conventions. Naive raters of the images produced by (esoteric) triads were 1.31 times as likely to be correctly identified as those of dyads.

In carrying out these experiments, there is also a demonstration of how such typological questions can be investigated experimentally (as suggested by Nettle, 2012). I have also illustrated how changes in linguistic complexity at either the level of the individual or in the shorter term, do not necessarily result in different degrees of complexity at the level of the group or in the longer term; additional mechanisms may be essential to solve the problem of linkage.

These studies have also added some support to other results in the literature. Experiments 1 and 3 replicate Saffran et al. (1996b) in illustrating that adult learners are able to use distributional cues to segment continuous linguistic input, and also extend the study to demonstrate their ability to generalise to novel speakers. The shorter training regimes of Experiment 3 also confirm Frank et al.'s (2010), admittedly unsurprising, findings that reduced exposure to such input will have a negative effect on performance. Experiment 8, meanwhile, adds further support to Kirby et al.'s (2015) conclusion that the emergence of language structure is reliant on learnability, rather than just expressivity, pressures.

While the investigations I have described have attempted to isolate and assess key components of the sociocultural effects proposed in Chapter 1, there are a number of claims which remain untested. Trudgill (2011) stresses that the features of communities which have more complex language — low amounts of adult language contact, high social stability, small size, dense social networks, and large amounts of communally-shared information — should be considered together, while I have deliberately isolated them. Interaction effects between the sociocultural factors have not been assessed. Trudgill (2011) also proposes the role of sound change in increasing the number of irregularities in a language or the number of morphological categories (see p. 89), and this is also untested.

Similarly, Lupyan and Dale's (2010) claims regarding redundancy aiding first language acquisition have not been assessed. I have concentrated on how esoteric communication may increase linguistic complexity, not how an increased pressure to be learnable by children over adults may have the same effect. There is also Wray and Grace's (2007) apparently competing hypothesis relating complexity to child language learning and processing. Wray and Grace (2007, p. 555) argue for some "default" level of holism in languages, in part "a product of the peculiar facility of the child to acquire language with recourse to full systematicity". For Lupyan and Dale (2010), internal structure appears to be what aids child learning: holistic items would not contain the syntagmatically redundant markers which may aid input segmentation, for example. It would be possible to test both of these claims experimentally.

5.3 Alternative determinants of linguistic complexity

I have so far focused on some of the key determinants which have been proposed to influence linguistic complexity, but other explanations have also been put forward. The next four sections discuss the proposed effects of language maturation, cultural complexity, group identity, and literacy.

5.3.1 Language maturation

One proposal is that "older" (or more "mature") languages are more complex. As argued by McWhorter (e.g. 2005, 2009), the simplest languages of the world are creoles,[54] and with time, such simple languages develop more complex features. Nichols (2009), however, notes that her analysis of linguistic complexity fails to provide much support for his initial claim that creoles are in fact simpler than other languages. Tok Pisin, one of the languages McWhorter (2005) refers to to illustrate creole simplicity, is the only creole in Nichols's (2009) sample. Though it ranks below average on her feature-based complexity measure, there are many non-creole languages which are simpler.[55]

The debate about creoles aside, we have already seen support for increases in complexity being a result of an increase in the amount of *mature phenomena*: language features which can only arise as the result of lengthy historical processes. Examples of such features include fusional morphology, agreement, obligatory grammatical markers, noun classes, and irregularities (Thurston, 1994; Dahl, 2004; McWhorter, 2005; Gil, 2009; Trudgill, 2011).[56] As we have seen, however, Trudgill (2011) argues that the conditions for such maturation are language acquisition and use in small, stable social groups with dense social networks, limited contact with other languages and in which large amounts of knowledge is shared.[57] It is not time, or the age of the language, which directly determines complexity. Determining a language's "age" to support such its relationship to certain language features is also controversial (DeGraff, 2001).

5.3.2 Group identity

As we have discussed, esoteric groups may create the ideal breeding ground for gradual language change and the development of more mature phenomena (Thurston, 1994; Dahl, 2004; Trudgill, 2011). This process may then be accelerated or exaggerated by the extent to which the speakers want to distinguish their language from others. Language can be used "as an in-group tool of expression and as an emblem of ethnic identity in contrast with other groups" (Thurston, 1994, p. 579-80). More esoteric communication may be the norm where groups wish to differentiate their language from neighbouring ones, which may be the case where neighbouring groups are disliked or where there are no natural barriers separating them. Communication may be less esoteric if there are natural barriers, such as mountain ranges or the sea, separating groups, or

[54] McWhorter (2005) illustrates this point using a feature-based complexity measure based on four components. Broadly, complexity is measured by the quantity of: marked members in a phoneme inventory ("those encountered less frequently in the world's languages than those conventionally deemed unmarked for example: ejectives, clicks, and labialized consonants vs. stops, rounded back vowels, and glides", McWhorter, 2005, p. 45); syntactic rules; overt grammaticalised distinctions equating to semantic or pragmatic distinctions; inflectional morphology.

[55] Also see McWhorter (2005, p. 69-71) and Gil (e.g. 2009) for a debate as to whether or not there exists an older, non-creole, language which creole-like simplicity. Gil (2009) proposes Riau Indonesian as such an example, while McWhorter (2005) argues that actually it should actually be considered a creole.

[56] See Dahl (2004) for a full account of the maturation processes involved. Also note that, depending on how the maturation process is described, that there is not necessarily agreement on what counts as a mature phenomenon. Bisang (2009), for example, disagrees with McWhorter's (2005) claims that linguistic tone is a mature feature.

[57] Proponents of the theory of the *morphological cycle*, that language change typically involves transition from isolating to agglutinating to fusional and back to isolating (Trudgill, 2011), may then argue that more complex language will not necessarily be found in such esoteric groups.

in the case of better relations between groups (Thurston, 1994).[58]

5.3.3 Cultural complexity

Perkins (1992) suggests that language complexity is inversely related to the complexity of the culture of a population. He considered a number of deictic markers on nouns and verbs related to nine variables of cultural complexity, relating to factors such as agricultural type and intensity, practised crafts and the extent of their specialisation, social class and organisational structures, city and population sizes, and property inheritance. He found a negative correlation between cultural complexity and the extent of deictic grammaticalisation.

Martowicz (2011) also found that the encoding of anteriority and conditionality is prone to the influence of sociocultural factors. Using a sample of 67 languages, she compared levels of written development, number of speakers and the presence and characteristics of language used in education, radio and television with the grammaticalisation, lexicalisation and explicitness of clause linkers. She therefore makes similar conclusions to Perkins (1992), that increased cultural complexity correlates with increased transparency and so lower levels of linguistic complexity.

Cultural complexity effects can also be seen when considering a language's lexicon. Thurston (1994, p. 600) gives the example of the languages spoken in north-western New Britain. The speakers of these languages had no traditions of offshore fishing or canoe building prior to the arrival of proto-Bariari speakers. As a result, "the lexicon covering this domain is copied wholesale from the nearest Bariari language". An increase in cultural complexity can lead to an increase in the size of the lexicon.

5.3.4 Literacy

As Maas (2009) and Martowicz (2011) discuss in detail, there may be substantial differences in the complexity of oral and literate language, both between languages and within the spoken and written form of the same language. The written language is very much a reshaped version of spoken languages, differing morphological and syntactically, using different vocabulary and organising a text in different ways. Crucially, the written form tends to be syntactically more complex. In English, for example, the written text uses a greater proportion of subordinate constructions, while spoken text uses a greater proportion of coordinate constructions. Unable to make use of extralinguistic cues, such as intonation, pitch, and speech rate (Chafe, 1987, from Martowicz, 2011), it needs to be more explicit than spoken language. This results in more complex syntax. There is also evidence that exposure to literate forms of a language increase a speaker's use of more complex syntactic constructions, and that languages with literate forms are more complex (Maas, 2009; Martowicz, 2011). Maas (2009, p. 177), for example, argues that "the world's simplest grammars are orate grammars".

Wray and Grace (2007), however, argue that literacy can lead to simplification. In forcing a greater level of analysis, a literate form of a language may lead to a greater amount of pattern recognition and application, such as found in more compositionally structured, and therefore simpler, language. Literacy may also maintain established simplicity: spaces between words make fusion less likely, and in doing so reduce a language's ability to become more complex (Deutscher, 2005). Experimental support can be found in Tamariz et al. (2010). Using

[58] Also see Roberts (2010) for a related experiment investigation into the effects of groups having a competitive motivation for differentiating their communication systems.

music and musical literacy as a proxy for language and literacy to investigate the effect of literacy on language processing, they find that musical literacy aids the retention of compositional structure. They suggest that this result would hold for language as well.

5.3.5 Interdependence effects

The identification of determinants of language can also be affected by interdependence, both within a set of candidate determinants of linguistic features, and within the set of linguistic features themselves. Interdependence can make it appear as though one non-linguistic factor is having a direct influence on language features when in fact it is not. As we saw in Chapter 3, for example, though total number of speakers is a proposed determinant of linguistic complexity, it is also correlated with proportion of non-native speakers (Lupyan and Dale, 2010). It may be that the proportion of non-native speakers is the better, or more directly influential, candidate determinant (Nettle, 2012). Similarly, terrain may influence the degree of esoteric communication. Mountainous landscapes, for example, may reduce the potential for communication with out-group members (Thurston, 1994). There is also the possibility of interaction effects. Smaller populations with no adult learning may have more complex language than larger populations with no adult learning, for example. As Trudgill (2011) argues, it may not always be appropriate to consider a set of possible determining factors in isolation. In Section 1.4, I also discussed the effects of the natural environment, biology, society and culture separately, but this is not intended to suggest that they are independent either. Environment influencing biology is one of the fundamental components of evolution, while biology influencing society and culture is also arguably trivial (Dediu, 2011). Examples such as the development of lactose tolerance in adults can be cited as evidence for cultural practice having the potential to guide genetic evolution. Indeed, in the development of phenomena such as language and the biological structure of the brain, there is the possibility of a process of coevolution, with each having an influence on the selective processes of the other (Christiansen and Chater, 2008). Environment has been argued to directly affect culture, with climate argued to determine the amount of expressivity in a social group, for example (Ember and Ember, 2007a), while clearly, society, culture and biology can in turn be seen to influence the environment in which they exist.

Individual linguistic features are also far from necessarily independent. There is a payoff between word length and phoneme inventory size (Trudgill, 2004a; Selten and Warglien, 2007; Sinnemäki, 2009), for example, while the existence of complex tone in a language is likely to diminish the functional load of intonation (Torreira et al., 2014), and increased morphology may increase the number of irregularities (Jackendoff, 1999). Language features, including those considered indicators of linguistic complexity, may be the result of other language features rather than non-linguistic determinants more directly. A “complex” language feature may itself increase complexity elsewhere. Alternations, for example, may increase paradigmatic redundancy (Trudgill, 2011, and see p. 89).

5.3.6 Exceptions

It is also worth noting that in identifying determinants of language features there will also be exceptions: we are of course “dealing with likelihoods, and formulating tendencies rather than strict rules of correlation” (Trudgill, 2011, p. ix). Therefore even if languages with more speakers are more complex, we may except to find some counterexamples. In some cases, it may be possible to propose explanations. Nichols (2009), for example, notes that a given population may have an unpredicted degree of language complexity due to a time lag between a change in social structure, and a slower change in linguistic structure. For example, smallpox epidemics in the Americas and Australia following European contact dramatically reduced population sizes, but have not yet significantly affected the complexity of their languages. Similarly, Lupyan and Dale (2010) suggest that political factors, such as the prescriptivism of twentieth century Russia, may counteract a tendency for a language's complexity to reduce as it spreads and the proportion of its adult learners increases.

5.4 Group size and complexity in other domains

Interestingly, the negative correlation between group size and linguistic complexity (Lupyan and Dale, 2010) appears to be at odds with two other proposed correlations: group size and the complexity of other cultural traditions in humans, and group size and the complexity of the communication systems of other species.

Larger groups have been proposed to result in increased cultural complexity in humans, with evidence from ethnographic research (Bell, 2015) and experimental studies (Derex et al., 2013; Muthukrishna et al., 2013; Kempe and Mesoudi, 2014). As discussed in Section 5.3.3 above, specific links between increased (non-linguistic) cultural complexity and decreased linguistic complexity have also been made (Perkins, 1992; Martowicz, 2011). We may be able to explain this discrepancy by considering the different *functions* of language and other cultural traditions. Language is a particular important cultural tradition, both in its ubiquity (Christiansen and Kirby, 2003) and in underpinning many other components of culture (Brighton et al., 2005; Sampson, 2009). As much of the complexity of language is arguably functionless (Dahl, 2004; Gil, 2009), and more simple, compositional languages have the potential to be generalisable (Kirby et al., 2008) and so be more expressive, there may be a functional benefit in language simplicity which may not necessarily be the case in other cultural domains.[59] Simpler language may allow increased complexity in the cultural traditions which are based on it.

Human language aside, communication systems generally appear to be more complex in larger or more complex social groups, such as those of non-human primates, sciurids, and bats (Nettle, 2012). The proposal that the social complexity of a group correlates with the complexity of their communication system is an idea that goes back as far as Lamarck and Darwin (Freeberg et al., 2012). This is based on the assumption that the more individuals have to interact with different individuals and of differing levels of hierarchical social structure, the greater the need for discrimination of communicative signals, or for a broader range of signals necessary to convey more complex information. The communicative system would therefore necessarily be more complex, and so, argue Freeberg et al. (2012), social complexity is the causal factor in such cases. They offer six predictors of communicative complexity with a group: size, density, member roles, egalitarian roles, size of home-range and stability. While their predictors show a great deal of similarity to Trudgill's (2011) determinants of linguistic complexity (see p. 15), they predict the opposite effect: increases in social complexity increases (non-linguistic, non-human) communicative complexity. Freeberg et al. (2012) go on to speculate that increased social complexity may well have been a factor in increasing the complexity of

[59] This does not rule out the possible functional benefits of higher degrees of complexity I have discussed elsewhere, such as the possible benefits for child learning (Lupyan and Dale, 2010) or defining group membership (Thurston, 1994).

pre-linguistic communication systems at earlier stages of our species' development, but following the emergence of language, the effect of social complexity has been very different. Human languages are also, after all, substantially more complex than the communication systems of all other species (Gil, 2009), are social learned, and are unboundedly expressive using combinations of elements from finite sets (Christiansen and Chater, 2008; Sampson, 2009; Trudgill, 2011; Nettle, 2012). They may also be unique amongst other communication systems in being subject to drift to some extent (Freeberg et al., 2012).

5.5 Directions for future research

There are a number of directions for future research suggested by this thesis, relating either immediately to the experiments, or to the field of non-linguistic determination of linguistic complexity more broadly.

The studies described here have demonstrated that hypotheses concerning sociocultural determination of language features can be experimentally investigated, and there are a number of possible follow ups. As I conclude in Atkinson et al. (2015) in Section 2.2, I would welcome replication of the input variability experiments, although ultimately believe that input variability is not a mechanism by which number of speakers could determine morphological complexity. That said, it would be interesting to scale up Experiment 1 to consider input variability affects on a much larger set of words. Frank et al. (2013) illustrated that adult learners are able to learn a language when the input was made up of 1,000 words types over 60,000 tokens, making up approximately 10 hours of data. This result could be replicated with a second condition in which the input was presented by multiple speakers. Given the amount of input, it would also be very possible to (greatly) increase the number of speakers in the high-variability condition from the 3 I used in Experiment 1.

There are some remaining questions from Chapter 3 which could also be investigated further. Firstly, it would be good to confirm adult learner simplification of morphology using a target language with less of a functional disparity between the stems and the suffixes than that of Experiment 5. Secondly, the possible linkage mechanisms — subsequent learning and native speaker accommodation — assessed in the rest of the chapter each leave a suggestion as to how individual-level simplification effects could propagate. Experiment 6 gives some suggestion that adult learner acquisition of a generalised subset of a target language could reduce group-level language complexity. A comparison of the Small-AllNative and Small-HalfNative conditions of Experiment 6 in which the simple input is such a variant of the complex could assess this directly. This could be supported by extending Experiment 5 to test different types of target language and considering second language acquisition research more widely to see if generalised subsets, rather than other simplified systems, of a language are more typical of the acquired morphological systems of adult learners. Experiment 6 could also be adapted to include social information about the speakers who provide the input. The complexity of the languages acquired by the participants may be different if each piece of the input includes some marker of speaker identity, indicating to the learner that any variation in their input is between-speaker, rather than within-speaker, variation.

Experiment 7 could also be adapted to focus on the effects of native speaker accommodation to non-native speakers. It would be particularly interesting to better understand the contexts in which native speakers are and are not prepared to accommodate, and under what circumstances they are prepared to go beyond simplifying their languages within "correct" grammatical grounds to produce ungrammatical utterances to aid non-native comprehension. Such an investigation could also be expanded to consider the effect of the interaction between children and non-native adult speakers, to see if child native speakers are more or less likely to accommodate than adults, or more or less likely to internalise any simplifications they use in such interactions. Ideally, this would be supported by a greater amount of research into cross-linguistic and cross-cultural foreign-directed speech.

Another possibility is to extend the experiments of Chapter 4 to more closely consider the effects of esoteric communication on the complexification of language. The group size and density manipulations of Experiment 8 could be combined with population turnover as in Kirby et al. (2015). The expectation would be the emergence of structure in the languages due to the increased pressure to be learnable, but the degree of that structure may turn out to be condition dependent. Again, a greater contrast between the experimental conditions would be recommended. It would also be worth seeing if the different degrees of description transparency for different size groups, as indicated by Experiment 11, is reproducible. Experiment 10 could also be adapted to consider different types of "shared knowledge". I investigated the effect of the distractor images being unique to each member of the group, but the images which made up the set for description could be manipulated so that they are (perhaps subtly) different for each speaker, for example. It would also be useful to develop alternative means of assessing esoteric effects on the semantic complexity of the emergent conventions, either through a more fine-grained analysis than I employed using the WordNet database, or changes to the experimental design.

Broadening the focus of possible future research, an investigation into how more complex language features could benefit child language acquisition (Wray and Grace, 2007; Lupyan and Dale, 2010, and see Section 5.2) could very feasibly be carried out using an artificial language learning paradigm. The claims of sociocultural determination of linguistic complexity (see Sections 1.5 and 5.3) can also be revisited and assessed. The effect of literacy (Wray and Grace, 2007) on the acquisition of compositional structure or redundant morphological marking could be evaluated experimentally, for example, while alternative possible determinants could also be proposed and considered. After our considerations of number of speakers (Lupyan and Dale, 2010), cultural complexity (Perkins, 1992; Martowicz, 2011), and features of the natural environment (Thurston, 1994; Trudgill, 2011), it may seem reasonable that a factor such as the degree of urbanisation in the areas a language is spoken in, for example, could influence language structure. More urban environments may be more likely to create the types of communities Trudgill (2011) argues would speak less complex languages (see p. 15). Similarly, if Foreigner Directed Speech does, as I suggest in Chapter 3, have implications for some linguistic features, it may also affect others. If native speaker accommodation to adult learners reduces structural complexity through the production of utterances with simplified grammar, then the hyperarticulation of the vowel space may also have some effect: languages with higher proportions of non-native speakers may have more sparse vowel spaces. Finally, it should be possible to assess any sociocultural effects on Bisang's (2009) notion of hidden complexity (see Section 1.3.4). This additional complexity which is manifested by additional inference necessary on the part of the hearer could possibly be assessed by measuring hearer processing or reaction times. If two signals have equal overt complexity, but one involves a greater amount of hearer processing (time), then it could be argued to have a greater degree of hidden complexity, for example.

Conclusion

The experimental evidence in this thesis ultimately supports the view that sociocultural factors can influence cross-linguistic variation in language complexity. I conclude that languages with a greater number of speakers may, as a result of their having a greater proportion of non-native speakers, be under a greater pressure to simplify. Languages of smaller groups, however, are more likely to maintain complexity and develop less transparent lexical items.

I investigated two candidate mechanisms by which number of speakers could determine morphological complexity: speaker input variability and adult learning. There is no evidence of an effect of speaker input variability: in two experiments, receiving linguistic input from a greater number of speakers had no effect on the segmentation of a continuous speech stream; and in two other experiments, on the acquisition of a morphologically complex miniature language.

Adult language learning is a plausible mechanism, however, assuming that languages with a greater number of speakers also have a greater proportion of non-native speakers. The first of three experiments found that adult learners do indeed simplify a morphologically complex miniature language. I then considered the problem of linkage: how these individual-level simplifications may affect language characteristics at the level of the group. Mixing the first experiment's simplifications with more complex input for a second generation of learners led to their receiving linguistic input which was in itself complex and variable, and this resulted in the simplifications being nullified. Language learning from input which includes the idiosyncractic simplifications of individual adult learners therefore cannot solve the problem of linkage on its own. The final experiment illustrates how language use, in the form of the native speaker accommodation to non-native speakers, may aid the propagation of adult learner simplifications, however. Participants trained on a language comprised of both regular and irregular verbs simplified their language to facilitate communication with a partner who was only trained on the regulars. Such accommodation may then increase the frequency and saliency of specific simplifications in the input of subsequent learners, leading to their propagation. Native speaker accommodation to non-natives may therefore be a key linking mechanism for adult learning leading to language-level morphological simplification.

There is also support for the maintenance of complexity in languages primarily used for esoteric communication, due to there being less of a pressure for learnability. In the absence of new learners, initially random artificial languages do not become more compositionally structured to a greater extent in larger or less dense social groups.

There is very little evidence for complexity *increasing* as a result of more esoteric communication, however. In two experiments which evaluated the conventionalisation of referring expressions for novel stimuli, there was no evidence of group size or quantity of shared information affecting the lengths or semantic complexity of the conventions, nor the extent to which they were systematically structured relative to the meaning space. In a final experiment in which naive individuals were required to match the conventions to their referents, however, there was evidence of an effect of group-size on the transparency of the referring expressions. Descriptions which developed in the larger groups were more accurately matched to their targets. There is therefore some indication that more esoteric communication may lead to the emergence of less transparent conventions from the point of view of out-group members.

Bibliography

Ansaldo, U., Lai, J., Jia, F., Siok, W. T., Tan, L. H., and Matthews, S. 2015. Neural basis for processing hidden complexity indexed by small and finite clauses in Mandarin Chinese. *Journal of Neurolinguistics*, 33:118–127.

Atkinson, M., Kirby, S., and Smith, K. 2015. Speaker input variability does not explain why larger populations have simpler languages. *PLoS ONE*, 10(6):e0129463.

Atkinson, Q. D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332(6027):346–9.

Ay, N., Müller, M., and Szkoła, A. 2008. Effective complexity and its relation to logical depth. *IEEE Transactions on Information Theory*, 56(9):4593–4607.

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.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Barton, M. E. and Tomasello, M. 1994. The rest of the family: the role of fathers and siblings in early language development. In Galloway, C. and Richards, B. J., editors, *Input and Interaction in Language Acquisition*, pages 109–134. Cambridge University Press, Cambridge.

Bates, D., Maechler, M., and Bolker, B. 2013. *lme4: Linear mixed-effects models using S4 classes*. Available at http://cran.r-project.org/package=lme4.

Bates, T. C., Luciano, M., Lind, P. A., Wright, M. J., Montgomery, G. W., and Martin, N. G. 2008. Recently-derived variants of brain-size genes ASPM, MCPH1, CDK5RAP and BRCA1 not associated with general cognition, reading or language. *Intelligence*, 36(6):689–693.

Beckner, C., Ellis, N. C., Blythe, R., Holland, J., Bybee, J., Christiansen, M. H., Larsen-freeman, D., Croft, W., and Schoenemann, T. 2009. Language is a complex adaptive system: Position paper. *Language Learning*, 59(Suppl. 1):1–26.

Bell, A. V. 2015. Linking observed learning patterns to the evolution of cultural complexity. *Current Anthropology*, 56(2):277–281.

Benders, T. 2013. Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, 36(4):847–862.

Bentz, C. and Winter, B. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, 3:1–27.

Bisang, W. 2009. On the evolution of complexity: sometimes less is more in East and mainland Southeast Asia. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 34–49. Oxford University Press, Oxford.

Bley-Vroman, R. W. 1989. What is the logical problem of foreign language learning? In Gass, S., editor, *Linguistic Perspectives on Second Language Learning*, pages 41–68. Cambridge University Press, Cambridge.

Boyd, R. and Richerson, P. J. 1985. *Culture and the Evolutionary Process.* University of Chicago Press, Chicago.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., and Tohkura, Y. 1999. Training Japanese listeners to identify English /r/ and /l/: long-term retention of learning in perception and production. *The Journal of the Acoustical Society of America*, 61(5):977–985.

Brainard, D. H. 1997. The Psychophysics Toolbox. *Spatial Vision*, 10:433–436.

Brighton, H. 2002. Compositional syntax from cultural transmission. *Artificial life*, 8(1):25–54.

Brighton, H. 2003. *Simplicity as a driving force in linguistic evolution.* PhD thesis, University of Edinburgh.

Brighton, H., Smith, K., and Kirby, S. 2005. Language as an evolutionary system. *Physics of Life Reviews*, 2(3):177–226.

Bromham, L., Hua, X., Fitzpatrick, T. G., and Greenhill, S. J. 2015. Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, page 201419704.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. 2002. What's new, pussycat? On talking to babies and animals. *Science*, 296(5572):1435.

Bybee, J. 2011. The vanishing phonemes debate, apropos of Atkinson 2011: How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. *Linguistic Typology*, 15:147–153.

Caldwell, C. A. and Smith, K. 2012. Cultural evolution and perpetuation of arbitrary communicative conventions in experimental microsocieties. *PLoS ONE*, 7(8):e43807.

Chafe, W. 1987. Cognitive constraints on information flow. In Tomlin, R., editor, *Coherence and Grounding in Discourse*, pages 21–51. Benjamins, Amsterdam.

Chater, N. and Vitányi, P. 2003. Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1):19–22.

Chipere, N. 2009. Individual differences in processing complex grammatical structures. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 178–191. Oxford University Press, Oxford.

Christiansen, M. H. and Chater, N. 2008. Language as shaped by the brain. *Behavioral and brain sciences*, 31(5):489–508; discussion 509–58.

Christiansen, M. H. and Kirby, S. 2003. Language evolution: the hardest problem in science? In Christiansen, M. H. and Kirby, S., editors, *Language evolution*, pages 1–15. Oxford University Press, New York.

Clark, E. V. 2003. *First Language Acquisition.* Cambridge University Press, Cambridge.

Clark, H. H. and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

Clark, R. 2001. Information theory, complexity and linguistic descriptions. In Bertolo, S., editor, *Language acquisition and learnability*, pages 126–171. Cambridge University Press, Cambridge.

Croft, W. 1995. Autonomy and functionalist linguistics. *Linguistic Soceity of America*, 71(3):490–532.

Croft, W. 2000. *Explaining language change: an evolutionary approach.* Longman, London.

Crutchfield, J. and Young, K. 1989. Inferring statistical complexity. *Physical review letters*, 63(2):105–108.

Crutchfield, J. P. and Whalen, S. 2012. Structural drift: the population dynamics of sequential learning. *PLoS computational biology*, 8(6):e1002510.

Csizér, K. and Dörnyei, Z. 2005. The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal*, 89(1):19–32.

Culbertson, J. and Newport, E. L. 2015. Harmonic biases in child learners: In support of language universals. *Cognition*, 139:71–82.

Cuskley, C., Colaiori, F., Castellano, C., Loreto, V., Pugliese, M., and Tria, F. 2015. The adoption of linguistic rules in native and non-native speakers: Evidence from a Wug task. *Journal of Memory and Language*, 84:205–223.

Dahl, Ö. 2004. *The Growth and Maintenance of Linguistic Complexity.* John Benjamins, Amsterdam.

Dahl, Ö. 2009. Testing the assumption of complexity invariance: the case of Elfdalian and Swedish. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 50–63. Oxford University Press, Oxford.

Dahl, Ö. 2011. Are small languages more or less complex than big ones? *Linguistic Typology*, 15(2):171–175.

Dale, R. and Lupyan, G. 2012. Understanding the origins of morphological diversity: the Linguistic Niche Hypothesis. *Advances in Complex Systems*, 15:1150017.

Dediu, D. 2011. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B, Biological Sciences*, 278(1704):474–9.

Dediu, D. and Ladd, D. R. 2007. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences of the United States of America*, 104(26):10944–10949.

DeGraff, M. 2001. On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics. *Linguistic Typology*, 5(1863):213–310.

Derex, M., Beugin, M.-P., Godelle, B., and Raymond, M. 2013. Experimental evidence for the influence of group size on cultural complexity. *Nature*, 503(7476):389–91.

Deutscher, G. 2005. *The unfolding of language.* Arrow books, London.

Deutscher, G. 2009. "Overall complexity": a wild goose chase? In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 243–251. Oxford University Press, Oxford.

Ember, C. R. and Ember, M. 2007a. Climate, econiche, and sexuality: Influences on sonority in language. *American Anthropologist*, 109(1):180–185.

Ember, C. R. and Ember, M. 2007b. Rejoiner to Munroe and Fought's commentary. *American Anthropologist*, 109(4):785.

Evans, N. and Levinson, S. C. 2009. The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448; discussion 448–492.

Everett, C. 2013. Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PLoS ONE*, 8(6).

Everett, C., Blasi, D. E., and Roberts, S. G. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, Early Edit:1–6.

Fay, N., Garrod, S., and Roberts, L. 2008. The fitness and functionality of culturally evolved communication systems. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 363(1509):3553–61.

Fehér, O., Kirby, S., and Smith, K. 2014. Social influences on the regularization of unpredictable variation. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 2187–2191, Austin, TX. Cognitive Science Society.

Ferguson, C. A. 1971. Absence of copula and the notion of simplicity: a study of normal speech, baby talk, foreigner talk and pidgins. In Hymes, D., editor, *Pidginization and Creolization of Languages*, pages 141–150. Cambridge University Press, Cambridge.

Ferguson, C. A. 1975. Trustees of Indiana University Anthropological Linguistics. *Anthropological Linguistics*, 17(6):305–323.

Fought, J. G., Munroe, R. L., Fought, C. R., and Good, E. M. 2004. Sonority and climate in a world sample of languages. *Cross-Cultural Research*, 38:27–51.

Frank, M. C., Goldwater, S., Griffiths, T. L., and Tenenbaum, J. B. 2010. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–25.

Frank, M. C., Tenenbaum, J. B., and Gibson, E. 2013. Learning and long-term retention of large-scale artificial languages. *PloS ONE*, 8(1):e52500.

Freeberg, T. M., Dunbar, R. I. M., and Ord, T. J. 2012. Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B, Biological sciences*, 367(1597):1785–801.

Frost, P. 2008. The spread of alphabetical writing may have favored the latest variant of the ASPM gene. *Medical Hypotheses*, 70(1):17–20.

Futrell, R., Mahowald, K., and Gibson, E. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 2015:201502134.

Garrod, S., Fay, N., Lee, J., Oberlander, J., and Macleod, T. 2007. Foundations of representation: where might graphical symbol systems come from? *Cognitive science*, 31(6):961–987.

Gell-Mann, M. 1994. *The Quark and the Jaguar: Adventures in the Simple and the Complex.* Little Brown, London.

Gil, D. 2009. How much grammar does it take to sail a boat? In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 19–33. Oxford University Press, Oxford.

Giles, H., Coupland, N., and Coupland, J. 1991. Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, J., and Coupland, N., editors, *Contexts of accommodation*, pages 1–68. Cambridge University Press, Cambridge.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Graf Estes, K. and Hurley, K. 2013. Infant-directed prosody helps infants map sounds to meanings. *Infancy*, 18(5):797–824.

Haspelmath, M. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Hay, J. and Bauer, L. 2007. Phoneme inventory size and population size. *Language*, 83(2):388–400.

Hockett, C. F. 1958. *A Course in Modern Linguistics.* Macmillan, New York.

Hudson Kam, C. L. and Newport, E. L. 2009. Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology*, 59(1):30–66.

Hurford, J. R. 2012. *The Origins of Grammar.* Oxford University Press, Oxford.

Iggesen, O. A. 2013. Number of Cases. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online.* Max Planck Institute for Evolutionary Anthropology.

Jackendoff, R. 1999. Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, 3(7):272–279.

Järvikivi, J., Vainio, M., and Aalto, D. 2010. Real-time correlates of phonological quantity reveal unity of tonal and non-tonal languages. *PLoS ONE*, 5(9):e12603.

Johnson, J. S. and Newport, E. L. 1989. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21:60–99.

Kaye, J. 1989. *Phonology: A Cognitive Review.* Erlbaum, Hillsdale, NJ.

Kempe, M. and Mesoudi, A. 2014. An experimental demonstration of the effect of group size on cultural accumulation. *Evolution and Human Behavior*, 35(4):285–290.

Kirby, S. 1999. *Function Selection and Innateness: the Emergence of Language Universals.* Oxford University Press, Oxford.

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):10681–6.

Kirby, S., Tamariz, M., Cornish, H., and Smith, K. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Kleiner, M., Brainard, D., and Pelli, D. 2007. What's new in Psychtoolbox-3? *Perception*, 36:ECVP Abstract Supplement.

Krause, J., Ruxton, G. D., and Krause, S. 2010. Swarm intelligence in animals and humans. *Trends in Ecology and Evolution*, 25(1):28–34.

Krauss, R. M. and Weinheimer, S. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. *Psychonomic Science*, 1(1-12):113–114.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686.

Li, M. and Vitányi, P. 1997. *An introduction to Kolmogorov complexity and its applications.* Springer, New York, third edition.

Little, H. 2011. *Foreigner directed speech: its role in cultural transmission of language & the resulting effects on language typology.* PhD thesis.

Lively, S. E., Logan, J. S., and Pisoni, D. B. 1993. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94:1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., and Yamada, T. 1994. Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, 96(4):2076–2087.

Long, M. H. 1981. Questions in foreigner talk discourse. *Language Learning*, 31(1):135–157.

Lupyan, G. and Dale, R. 2010. Language structure is partly determined by social structure. *PLoS ONE*, 5(1):e8559.

Lytinen, S. L. 1987. Integrating syntax and semantics. In Nirenburg, S., editor, *Machine Translation.* Cambridge University Press, Cambridge.

Maas, U. 2009. Orality versus literacy as a dimension of complexity. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 164–177. Oxford University Press, Oxford.

Marten, K. and Marler, P. 1977. Sound transmission and its significance for animal vocalization II. Tropical forest habitats. *Behavioral Ecology and Sociobiology*, 2(3):291–302.

Martowicz, A. 2011. *The Origin and Functioning of Circumstantial Clause Linkers : A Cross-Linguistic Study.* PhD thesis, University of Edinburgh.

McAllister, J. 2003. Effective complexity as a measure of information content. *Philosophy of Science*, 70(2):302–307.

McWhorter, J. 2009. A bewilderingly multifunctional Saramaccan word teaches us how a creole language develops complexity. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 141–163. Oxford University Press, Oxford.

McWhorter, J. H. 2005. *Defining Creole.* Oxford University Press, Oxford.

Meir, I., Israel, A., Sandler, W., Paddon, C., and Aronoff, M. 2012. The influence of community on language structure: evidence from two young sign languages. *Linguistic Variation*, 12(2):247–291.

Mietsamo, M. 2009. Implicational hierarchies and grammatical complexities. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 80–97. Oxford University Press, Oxford.

Mills, G. J. and Healey, P. G. T. 2016. Submitted. A Dialogue Experimental Toolkit.

Munroe, R. L. and Fought, J. G. 2007. Response to Ember and Ember's "Climate, econiche, and sexuality: Influences on sonority in language". *American Anthropologist*, 109(4):784–785.

Munroe, R. L., Fought, J. G., and Macaulay, R. K. S. 2009. Warm climates and sonority classes: Not simply more vowels and fewer consonants. *Cross-Cultural Research*, 43(2):123–133.

Munroe, R. L. and Silander, M. 1999. Climate and the consonant-vowel (CV) syllable: A replication within language families. *Cross-Cultural Research*, 33:43–62.

Muthukrishna, M., Shulman, B. W., Vasilescu, V., and Henrich, J. 2013. Sociality influences cultural complexity. *Proceedings of the Royal Society B, Biological Sciences*, 281:20132511.

Nettle, D. 1999a. Is the rate of linguistic change constant? *Lingua*, 108(2-3):119–136.

Nettle, D. 1999b. *Linguistic Diversity.* Oxford University Press, Oxford.

Nettle, D. 2012. Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1597):1829–1836.

Newmeyer, F. J. 2002. Uniformitarian assumptions and language evolution research. In Wray, A., editor, *The Transition to Language*, pages 359–375. Oxford University Press, Oxford.

Newport, E. L. 1990. Maturational constraints on language learning. *Cognitive science*, 14:11–28.

Newport, E. L. 2002. Language Development, Critical Periods in. *Encyclopedia of Cognitive Science*, pages 737–740.

Nichols, J. 2009. Linguistic complexity: a comprehensive definition and survey. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 110–125. Oxford University Press, Oxford.

Peirce, J. W. 2007. PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2):8–13.

Pericliev, V. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology*, 8:376–383.

Perkins, R. 1992. *Deixis, Grammar and Culture.* Benjamins, Amsterdam.

Pine, J. M. 1994. The language of primary caregivers. In Galloway, C. and Richards, B. J., editors, *Input and Interaction in Language Acquisition*, pages 15–37. Cambridge University Press, Cambridge.

Pullum, G. K. 1991. *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language.* University of Chicago Press, Chicago.

R Core Team 2013. *R: A language and environment for statistical computing.* Available at http://www.r-project.org/.

Regier, T., Carstensen, A., and Kemp, C. 2016. Languages Support Efficient Communication About The Environment: Words For Snow Revisited. In Roberts, S. G., Cuskley, C., McCrohon, L., Barceló-Coblijn, L., Fehér, O., and Verhoef, T., editors, *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11).* Online at http://evolang.org/neworleans/papers/54.html.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry Theory.* World Scientific, London.

Roberts, G. 2010. An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction Studies*, 11(1):138–159.

Saffran, J. R., Aslin, R. N., and Newport, E. L. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Saffran, J. R., Newport, E. L., and Aslin, R. N. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 621(35):606–621.

Sampson, G. 2009. A linguistic axiom challenged. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 1–18. Oxford University Press, Oxford.

Sapir, E. 1912. Language and environment. *American Anthropologist*, 14(2):226–242.

Schreier, D. 2003. *Isolation and Language Change: Contemporary and Sociohistorical Evidence from Tristan da Cunha English.* Palgrace Macmillan, Basingstoke.

Scovel, T. 2000. A critcal review of the critical period research. *Annual Review of Applied Linguistics*, 20:213–223.

Selinker, L. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10:209–231.

Selten, R. and Warglien, M. 2007. The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7361–6.

Shannon, C. E. and Weaver, W. W. 1949. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Simons, A. M. 2004. Many wrongs: The advantage of group navigation. *Trends in Ecology and Evolution*, 19(9):453–5.

Sinnemäki, K. 2009. Complexity in core argument marking and population size. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 126–140. Oxford University Press, Oxford.

Sinnemäki, K. 2011. *Language universals and linguistic complexity Three case studies in core argument marking.*

Smith, K. 2012. Evolutionary perspectives on statistical learning. In Rebuschat, P. and Williams, J. N., editors, *Statistical Learning and Language Acquisition*, pages 409–432. De Gruyter Mouton, Berlin.

Smith, K. and Kirby, S. 2008. Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B, Biological Sciences*, 363(1509):3591–603.

Snedeker, J. and Trueswell, J. C. 2004. The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49(3):238–99.

Szmrecsany, B. and Kortmann, B. 2009. Between simplification and complexification: non-standard varieties of English around the world. In Sampson, G., Gil, D., and Trudgill, P., editors, *Language Complexity as an Evolving Variable*, pages 64–79. Oxford University Press, Oxford.

Tamariz, M., Brown, J. E., and Murray, K. M. 2010. The role of practice and literacy in the evolution of linguistic structure. In Smith, A. D. M., Schouwstra, M., de Boer, B., and Smith, K., editors, *Proceedings of the 8th International Conference on the Evolution of Language*, pages 313–320. World Scientific.

Thiessen, E. D., Hill, E. A., and Saffran, J. R. 2005. Infant-directed speech facilitates word segmentation. *Infancy*, 7(1):53–71.

Thurston, W. R. 1994. Renovation and innovation in the languages of north-western New Britain. In Dutton, T. and Tryon, D. T., editors, *Language Contact and Change in the Austronesian World*, pages 573–609. Mouton De Gruyter, Berlin.

Torreira, F., Roberts, S. G., and Hammarström, H. 2014. Functional trade-off between lexical tone and intonation: Typological evidence from polar-question marking. *Proceedings of the 4th International Symposium on Tonal Aspects of Languages (TAL 2014)*, pages 100–103.

Trudgill, P. 2004a. Linguistic and social typology : The Austronesian migrations and phoneme inventories. *Linguistic Typology*, 8:305–320.

Trudgill, P. 2004b. On the complexity of simplification. *Linguistic Typology*, 8:384–388.

Trudgill, P. 2010. Contact and sociolinguistic typology. *Handbook of Language Contact*, pages 299–319.

Trudgill, P. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity*. Oxford University Press, Oxford.

Trueswell, J. C., Sekerina, I., Hill, N. M., and Logrip, M. L. 1999. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition*, 73(2):89–134.

Uther, M., Knoll, M. A., and Burnham, D. 2007. Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, 49(1):2–7.

Weighall, A. R. 2008. The kindergarten path effect revisited: children's use of context in processing structural ambiguities. *Journal of experimental child psychology*, 99(2):75–95.

Wesche, M. B. 1994. Input and interaction in second language acquisition. In Gallaway, C. and Richards, B. J., editors, *Input and Interaction in Language Acquisition*, pages 219–250. Cambridge University Press, Cambridge.

WordNet 3.1 2010. About WordNet. In *WordNet*. Princeton University.

Wray, A. and Grace, G. W. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, 117(3):543–578.

Wright, L. 2012. On variation and change in London medieval mixed-language business documents. In Stenroos, M., Mäkinen, M., and Saerheim, I., editors, *Language Contact and Development Around the North Sea*, pages 99–118. John Benjamins, Amsterdam.

Wright, L. 2013. The contact origins of Standard English. In Schreier, D. and Hundt, M., editors, *English As a Contact Language*, pages 58–74. Cambridge University Press, Cambridge.

Xu, N., Burnham, D., Kitamura, C., and Vollmer-Conna, U. 2013. Vowel Hyperarticulation in Parrot-, Dog-and Infant-Directed Speech. *Antrozoös*, 26(3):373–380.

Xu Rattanasone, N., Burnham, D., and Reilly, R. G. 2013. Tone and vowel enhancement in Cantonese infant-directed speech at 3, 6, 9, and 12 months of age. *Journal of Phonetics*, 41(5):332–343.

A Linear mixed effects analyses

Linear mixed effects modelling has played some role in the analysis of each of the experiments in this thesis, and this section is intended to clarify the approach I have taken in each case and avoid unnecessary replication elsewhere.

Each analysis used R (R Core Team, 2013) and *lme4* (Bates et al., 2013). Following Barr et al. (2013), a maximal model was constructed, but constrained by theoretical motivations. Therefore all effects and their interactions which could potentially influence the dependent variable were included, but those with no justification were not. I have included some explanation of the effects included in the model where it seemed necessary (see, for example, the model on p. 42). Where I analyse binary data, the model employs logistic regression. Where fixed effects have been centred (to reduce collinearity effects), this is marked.

Each model was first compared to its null equivalent, which was a secondary model with the same random effect(s), but with all fixed effects removed. The results of this comparison are given in the text. A non-significant difference between the model and its null equivalent is interpreted as none of its fixed effects being significant. If there is a significant difference between the model and the null, and both the Akaike information criterion (AIC) and the Bayesian (BIC) are lower for the model than the null, the model is assumed to fit the data better and its effects analysed. In the cases where the AIC indicates a better fit of the data, but the BIC indicates overparameterisation, this is described and discussed in the text.

For each fixed effect which is indicated as accounting for a significant amount of the model variance, I give the β coefficient and standard error. With logistically fit models, I also give the p-values of the R output. Otherwise, p-values are estimated from the resultant t-statistics with degrees of freedom being the number of observations minus the number of fixed parameters in the model (Baayen et al., 2008). I give these t-statistics and degrees of freedom in the text. The same information is given for all of the non-significant fixed effects, although this information is often pooled for brevity (as, for example, on p. 42). Where the variables are compared to some baseline condition, that baseline is noted in the text.

A p-value less than or equal to 0.05 is taken to be statistically significant.

B Experiment 5 pilot

A preliminary investigation was carried out to test the appropriateness of the target language, stimuli, and general experimental procedure, and also to determine an appropriate number of rounds of training and testing to include in the design.

B.1 Materials and methods

The experimental design was the same as that described in Section 3.2.1, except the number of rounds was set to 6, rather than 8.

B.2 Participants

11 native English speakers (1 male; aged between 19 and 23, mean 20.9) were recruited at the University of Edinburgh. Each was compensated £7. The experiment was run between 12th and 15th May 2014. At they exited the experiment, it was clear that 2 of the participants (both female; aged 19 and 20) were considerably intoxicated, and so their data is not included in any analysis.

B.3 Analysis and results

Stem accuracy, suffix accuracy and suffix stability between rounds was calculated as in Section 3.2.3. The results for the accuracy measures are illustrated in Figure 53.

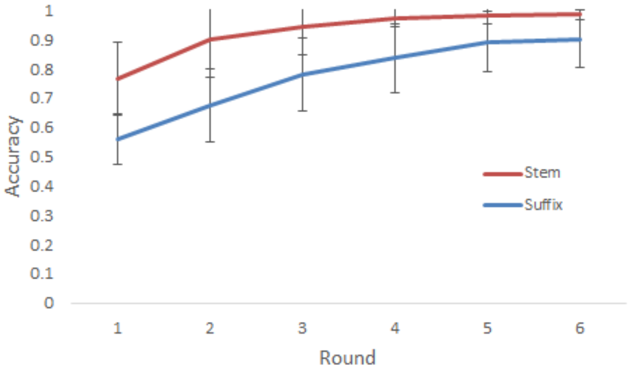


Figure 53: Average stem and suffix accuracy by round. Individual accuracy scores calculated as 1 minus the normalised Levenshtein distance between the production and the target. Error bars are 95% confidence intervals.

Both stem accuracy (Welch's two-sample t-test, t(8.328) = -4.04, p = 0.003) and suffix accuracy (t(15.757) = -5.943, p <0.001) were significantly greater in Round 6 compared to Round 1. Stem accuracy appears to be close to ceiling by Round 4. 7 of the 9 participants had a stem accuracy score ≥ 0.92 by Round 2, and 8 out of 9 had a score ≥ 0.96 by Round 3. Stability is illustrated in Figure 54, and is significantly greater at Round 6 than at Round 2 (t(11.571) = -5.314, p <0.001).

Figures 53 and 54 suggest that the suffixes used for each meaning were approaching those of the target language and a degree of stability by Round 5 or 6. Average suffix success is 0.89

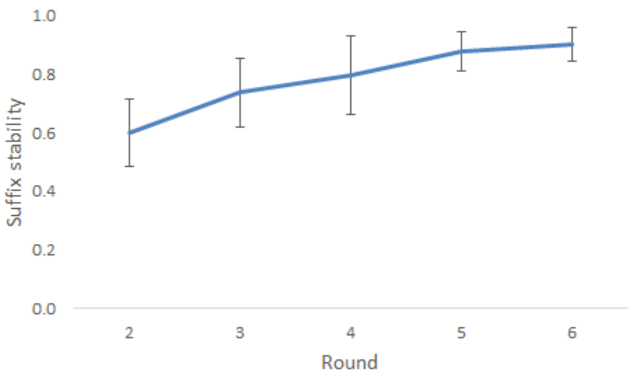


Figure 54: Suffix stability by round. Individual stability scores calculated as 1 minus the normalised Levenshtein distance between the produced suffix at that round and the previous production for the same stimulus. Error bars are 95% confidence intervals.

± 0.10 (95% CI) at Round 5 and 0.90 ± 0.10 at Round 6. Suffix stability is 0.88 ± 0.07 at Round 5 and 0.90 ± 0.06 at Round 6.

The experimental design seemed to be appropriate. The task was clear to the participants and the noun stems were quickly acquired while the suffixes proved more challenging. It was not completely clear whether suffix success and suffix stability would increase further if there had been a greater number of rounds, however. For the experiment proper, therefore, the number of rounds was increased to 8.

C Experiment 5 supplementary analysis

In addition to the complexity measures involving entropy and mutual information described in Section 3.2.3.2, I also considered 3 other measures, calculating statistical complexity, number of rewrite rules, and Mantel tests.

C.1 Statistical complexity

While the entropy measure considers the suffixes of each word independently, statistical complexity (e.g. Crutchfield and Young, 1989; Crutchfield and Whalen, 2012) is an alternative meaning-independent measure which considers the paths from Q to N to V. Statistical complexity is calculated as:

$$C(S) = -\sum_{\sigma \in S} P(\sigma) \log_2 P(\sigma) \tag{2}$$

where σ is a node on the possible paths.

For example, the nodes and paths for the "simple" language in Table 6 on p. 58 are shown in Figure 55. Here, $C(S) = 4.387$. By comparison, the "complex" , with the paths of Figure 56, has $C(S) = 6.881$.

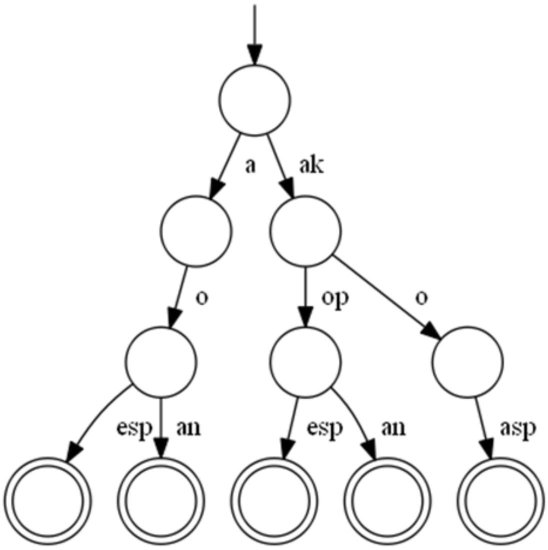


Figure 55: Paths for the "simple" example. See Table 6.

Average $C(S)$ by round with comparison to that of the target language is shown in Figure 57. The increase with round is significant (L=4502, p <0.001), but there is no evidence of a difference between $C(S)$ at Round 2 and at Round 8 (t(27.794) = 1.566, p = 0.129).[60] So

[60] There is also no evidence that $C(S)$ at either Round 2 (t(25) = 1.666, p = 0.108) or Round 8 (t(25 = -1.215, p = 0.236) is significantly different to that of the target language.

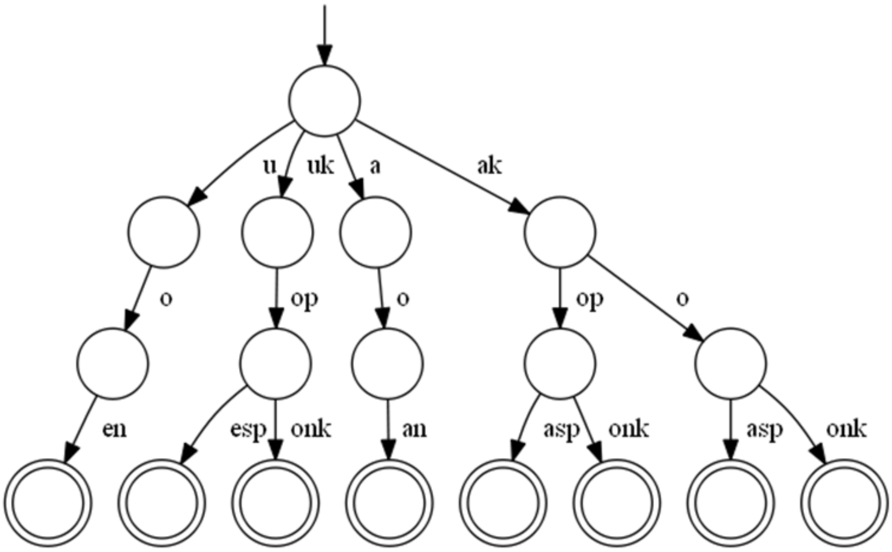


Figure 56: Paths for the "complex" example. See Table 6.

though we have some evidence that complexity increases with increased participant training and testing and so is generally lower in earlier rounds, there is no specific evidence that it is lower at Round 2 than at Round 8.

C.2 Rewrite rules

An alternative meaning-dependent way to quantify the complexity of the morphological systems is to consider the simplest set of rules which can exactly recreate the produced suffixes. We consider two measures: the minimum number of rules needed to describe the system, and the description length of those rules.

The rule-defining process is best described using an example. Consider the "simple" language's quantifier set (Table 6 on p. 58) and its relationship to the meaning space (Table 5 on p. 56). We begin with (one of) the majority suffixes as the default suffix, and then add more specific rules until the entire set is completely described:

Rule	Suffix	Condition
1.	a	
2.	ak	Number = 2

The set can then be reconstructed by applying the more specific rules first. In this case we would have "ak" if Number is 2, and "a" otherwise. The number of rules here is 2. The description length is then the number of suffixes plus the number of conditional statements, in this example 3. A less trivial example is the process for V for the "complex" set (Table 6 on p. 6):

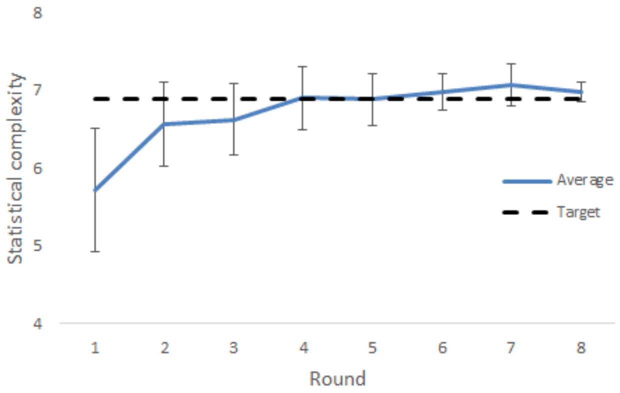


Figure 57: Average $C(S)$ by round. Statistical complexity of target language is 6.881. Error bars are 95% confidence intervals.

Rule	Suffix	Conditions		
1.	an			
2.	asp	Number = 2		
3.	en		Animal = crocodile	
4.	esp	Number = 2	Animal = crocodile	Movement $\neq$ V3
5.	onk	Number = 2		Movement = V3

Here the number of rules is 5, and the description length is 12.[61,62]

The total number of rules and description length for a given participant's productions are the total of those of Q, N and V. The average number of rules by round is illustrated in Figure 58, and the average description length by round in Figure 59.

For the number of rules measure, a one-way MANOVA was conducted with *round* (Round 2 or 8) as the independent variable and the *number of quantifier*, *noun* and *verb suffix rules* as dependent variables. There was a marginally significant effect of *round* (Pillai's trace= 0.145, F(3,48) = 2.721, p = 0.055). The description lengths were calculated in the same way, and also indicate a marginally significant effect of *round* (Pillai's trace= 0.123, F(3,48) = 2.247, p = 0.095).

Under these two measures, we therefore have only limited evidence of a difference in the complexity of Round 2 and Round 8 languages.

C.3 Mantel tests

For each of Q, N and V for each production set, Mantel tests were run to give some indication of how structured the suffixes were relative to the meaning space (as in, e.g., Kirby et al., 2008). The Hamming distance between all pairs of meanings, and the normalised Levenshtein

[61] To reconstruct the suffix set, apply more specific rules (i.e. the ones with a greater number of conditions) before less specific ones. Rule 4 (3 conditions) then gives the V suffixes for Stimuli 2 and 4 in the meaning space (see Table 6). Rule 5 (2 conditions) then labels Stimuli 6, 12 and 18. Rules 2 and 3 can then be applied in either order (both have 1 condition), and label Stimuli 1, 3 and 5, and 8, 10, 14 and 16, respectively. Rule 1 (0 conditions) is then applied to all remaining stimuli.

[62] At first glance, it may appear that the description length can be reduced, for example by removing the "Movement $\neq$ V3" condition from Rule 4. This would lead to ambiguity in the reconstruction process, however, as the V suffix for Stimulus 6 (2 crocodile V3) would be either "esp" or "onk", depending on which rule was applied first.

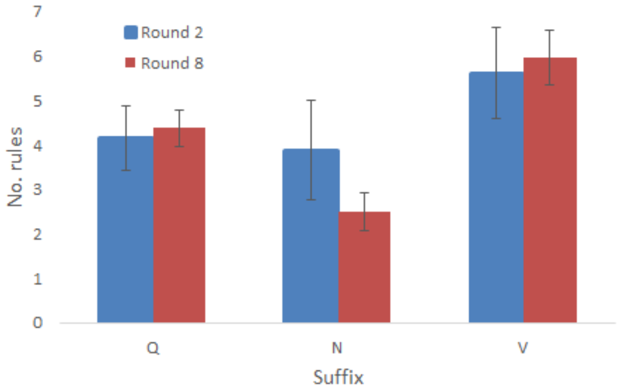


Figure 58: Average number of rules for Rounds 2 and 8. Error bars are 95% confidence intervals.

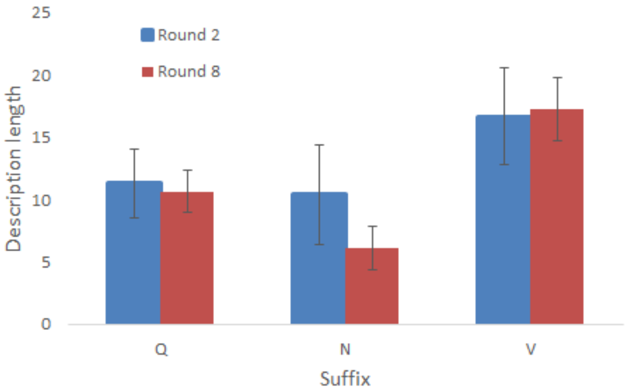


Figure 59: Average description length for Rounds 2 and 8. Error bars are 95% confidence intervals.

distance between all pairs of suffixes were calculated, and Pearson's product-moment correlation calculated for these two distance sets. To compare different languages, z-scores were calculated using Monte Carlo simulations with 1,000 randomisations. The average z-scores by round for Q, N and V are illustrated in Figures 60, 61 and 62, respectively.

A one-way MANOVA was conducted with *round* (Round 2 or 8) as the independent variable and the *quantifier*, *noun* and *verb suffix structure* as dependent variables. There was a significant effect of *round* (Pillai's trace= 0.232, F(3,48) = 4.8265), p = 0.005), on *quantifier suffix structure* (t(45.933) = -2.101, p = 0.041) and *verb suffix structure* (t(49.878) = -3.2289, p = 0.002), but not on *noun suffix structure* (t(48.227) = -0.053, p = 0.958).

There is evidence here then, that the suffixes become more structured relative to the meaning space with increased participant training and testing, at least for the quantifiers and verbs.

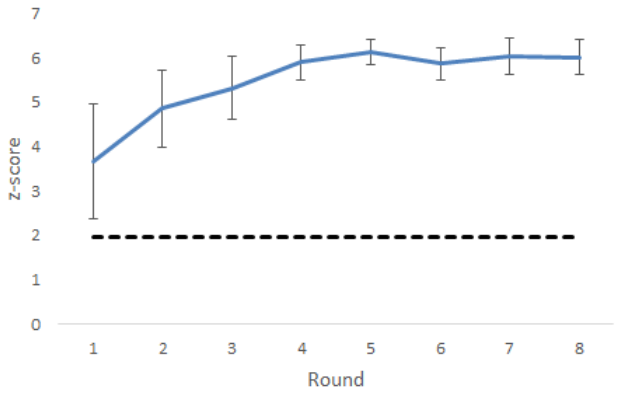


Figure 60: Z-scores for Q by round. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

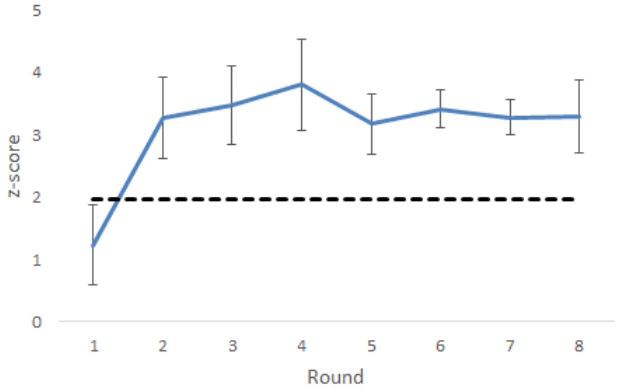


Figure 61: Z-scores for N by round. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

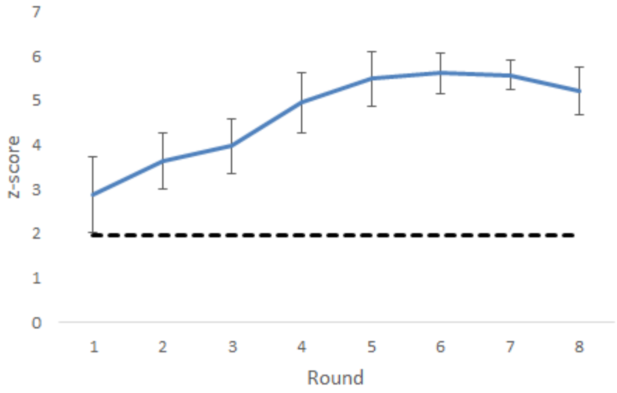


Figure 62: Z-scores for V by round. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

D Experiment 6 supplementary analysis

As for Experiment 5 (Appendix C), we also considered statistical complexity, number of rewrite rules and Mantel tests.

D.1 Statistical complexity

Average statistical complexity is show in Figure 63. There is no significant difference between the conditions for the final round productions (F(3,44) = 0.615, p = 0.609).

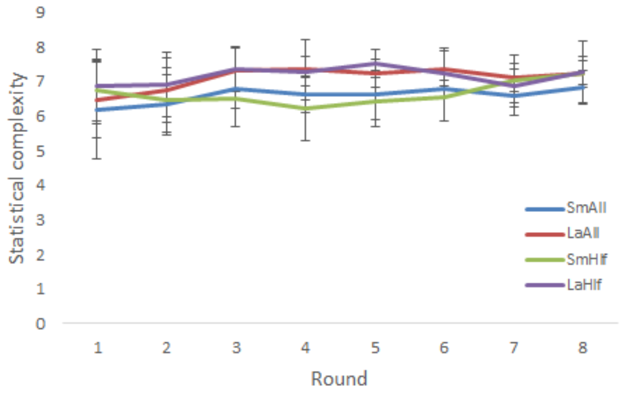


Figure 63: Average statistical information by round and condition. Error bars are 95% confidence intervals.

D.2 Rewrite rules

Average number of rules and descriptions lengths by condition are illustrated in Figures 64 and 65, respectively.

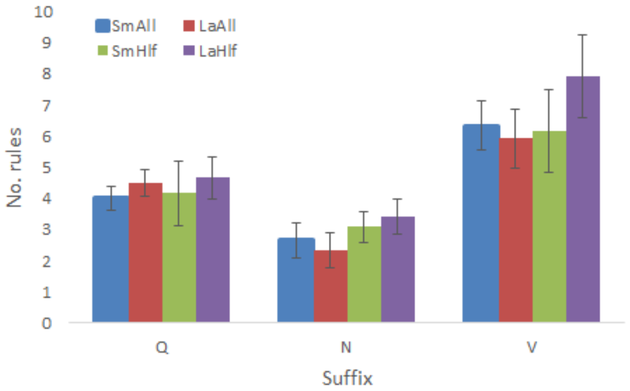


Figure 64: Average number of rules by condition. Error bars are 95% confidence intervals.

A one-way MANOVA indicated a significant effect of condition for the number of rules (Pillai's trace = 0.374, F(9,132) = 2.087, p = 0.035). One-way ANOVAs found a significant effect of *condition* on *number of noun rules* (F(3,44)=3.575, p = 0.021) and *number of verb*

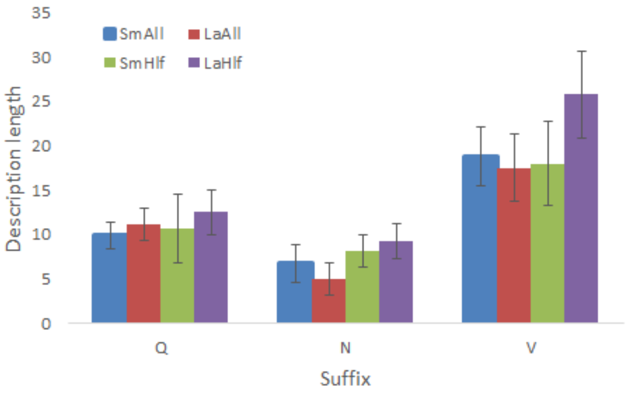


Figure 65: **Average description length by condition.** Error bars are 95% confidence intervals.

rules (F(3,44) = 3.178, p = 0.033), but no effect on *number of quantifier rules* (F(3,44) = 0.952, p = 0.424). Mean *number of noun rules* of LaHlf was greater than Large-AllNative (diff = 1.083, p = 0.019), but there was no significant difference between any of the other pairs of means (p ≥ 0.164). Mean *number of verb rules* for LaHlf was significantly greater than for Large-AllNative (diff = 2.000, p = 0.038). There was no significant difference between any of the other means (p ≥ 0.085).

There was also a significant effect of condition for the description lengths (Pillai's trace = 0.419, F(9,132) = 2.380, p = 0.016) on *noun description length* (F(3,44) = 4.491, p = 0.008) and *verb description length* (F(3,44) = 3.989, p = 0.013), but not *quantifier description length* (F(3,44) = 0.853, p = 0.473). Multiple comparison of means for the *noun description length* indicated that LaHlf was significantly greater than Large-AllNative (diff = 4.333, p = 0.006). No other mean differences were significant (p ≥ 0.068). For the means of *verb description length*, LaHlf was significantly greater than Large-AllNative (diff = 8.250, p = 0.021), and LaHlf was significantly greater than SmHlf (diff = 7.750, p = 0.033). No other differences were significant (p ≥ 0.069).

Again, we have only limited evidence of a difference between the conditions here, but there is some suggestion that for N and V, both the number of rules and the description lengths are greater for LaHlf than for Large-AllNative. This difference would imply that LaHlf languages are more complex than Large-AllNative, if anything. This would be at odds with the proposal that greater proportions of non-native speakers are correlated with less complex languages (e.g Lupyan and Dale, 2010).

D.3 Mantel tests

Average z-scores for each of Q, N and V by condition are shown in Tables 66, 67 and 68, respectively.

There is a significant effect of condition (Pillai's trace = 0.460, F(3,44) = 2.654, p = 0.007) on *verb structure*(F(9,44) = 5.528, p = 0.003), but no significant effects on *noun structure* (F(9,44) = 2.759, p = 0.053) or on *quantifier structure* (F(9,44) = 1.846, p = 0.153). For *verb structure*, Large-AllNative was significantly greater than LaHlf (diff = 1.767, p = 0.045), Small-AllNative was significantly greater than LaHlf (diff = 2.333, p = 0.005), and Small-

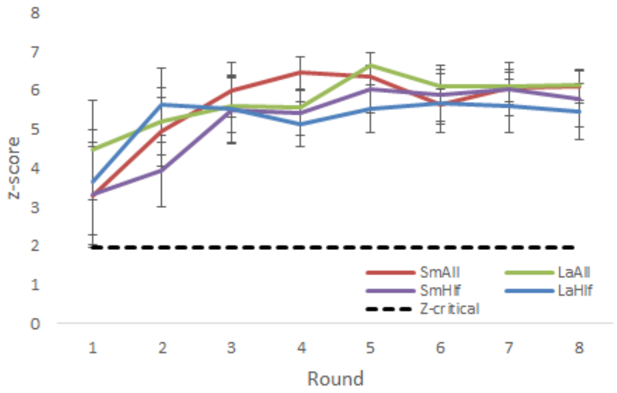


Figure 66: Mantel tests for Q by round and condition. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

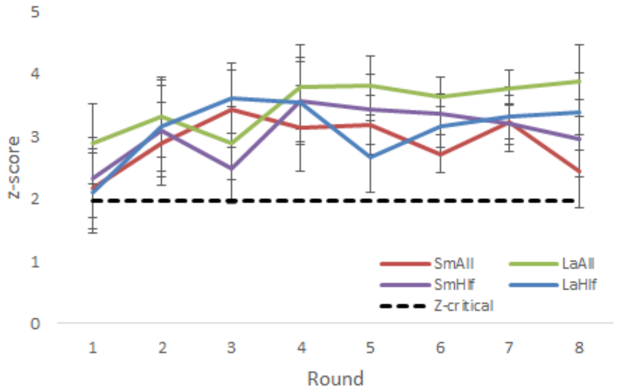


Figure 67: Mantel tests for N by round and condition. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

AllNative was significantly greater than SmHlf (diff = 1.831, p = 0.036). All other differences were non-significant ($p \geq 0.226$).

For the verb suffix then, there is some evidence that the All conditions are more structured with respect to the meaning space than the Hlf conditions, and therefore simpler, more transparent, and more regular.

D.4 Input statistical complexity

Another meaning-independent measure of the variability of the input is to calculate the statistical complexity of the 72 input strings received by each participant. The averages by condition are shown in Figure 69.

There is a significant difference between the conditions (F(3,44) = 26.2, p <0.001). LaHlf is significantly greater than Large-AllNative (diff = 1.083, p <0.001), LaHlf greater than Small-AllNative (diff = 1.329, p <0.001), LaHlf greater than SmHlf (diff = 0.796, p <0.001), and

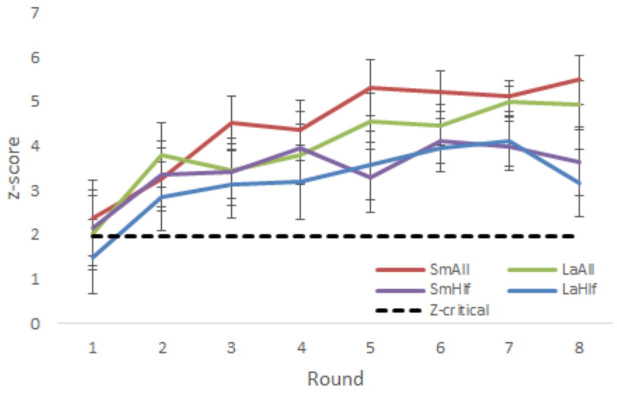


Figure 68: Mantel tests for V by round and condition. The dashed line indicates the critical value (z=1.96) at 95% confidence. Scores above this line imply significant structure of the signal space to the meaning space. Error bars are 95% confidence intervals.

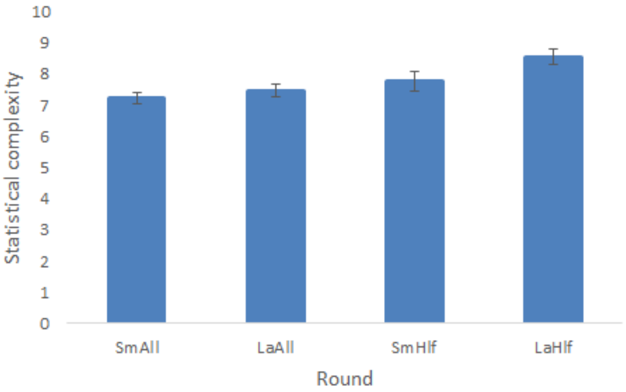


Figure 69: Input statistical complexity. Error bars are 95% confidence intervals.

SmHlf greater than Small-AllNative (diff = 0.533, p = 0.009). There was no significant difference between Small-AllNative and Large-AllNative (p = 0.419) and SmHlf and Large-AllNative (p = 0.288).

See Section D.5 for simulation results relating to this measure.

D.5 Simulation of the input

We can get a better idea of the variability of the input likely to arise in the different conditions through simulating the process of assigning speakers and input data. Figure 70 shows the statistical complexity of 10,000 randomly sampled inputs for each condition (compare to Section D.4 for the statistical complexity of the actual input of the experiment).

Two extra conditions are added to the simulation of inputs (marked in yellow in Figure 70): a small population (n=2) with all non-native speakers (SmZro), and a large population (n=8) with all non-native speakers. We see that the input of the Hlf conditions again appears to be more statistically complex than the All conditions, and that the LaHlf input is more statistically complex than the SmHlf. We can also see that the LaZro input is statistically complex than

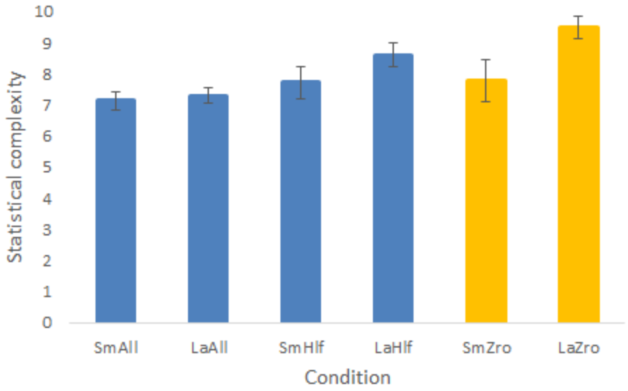


Figure 70: Statistical complexity of 10,000 simulated input datasets. Error bars are standard deviation.

any of the other conditions, and that SmZro is around as complex as SmHlf.

Clearly mixing simple languages with complex, or simple with simple, does not result in simpler input here than for more consistent input, even if the consistent input is of a relatively complex language.

E Experiments 9 and 10 Occurrence 4 labels

Table 11 gives the full list of descriptions assessed for transparency by naive raters, as described in Section 4.5.

Image	Group	Label
A01	dy02	the camel with one hump
A01	dy03	camel/ostrich with giant arrow body
A01	dy04	the camel with one hump
A01	es08	single hunchback camel looking left
A01	ex01	the camel (one hump)
A01	ex03	one hump camel
A01	ex09	camel, one hump, facing left
A01	ex11	one humped camel lefty
A02	dy08	The camel and 2 humps
A02	dy12	camel with two humps facing left
A02	es01	the camel
A02	es03	the camel
A02	es09	camel
A02	es11	the camel
A02	es12	the 2 cocoase camel
A02	ex03	camel two hump
A02	ex08	camel
A02	ex11	2 humped camel facing left
A03	es06	the turkey/ostrich
A03	es07	square triangle on the left with square triangle on the left
A03	ex02	spider with house head
A03	ex05	flying monkey
A04	dy06	wolf howling
A04	dy07	wolf howling
A04	es01	howling wolf
A04	es08	diamond dog eating food
A04	es12	the hand
A04	ex01	the wolf
A05	dy04	giraffe looking to left
A05	dy12	giraffe facing left
A05	es02	giraffe facing left
A05	es08	dog looking at left
A05	ex03	dog looking left
A06	es03	the dog
A06	ex09	dog, facing right
A06	ex10	dog/reclining chair
A07	dy01	not-dog facing right
A07	dy08	the lizard with long neck again facing right
A07	dy13	the llama thing with parallelogram first leg
A07	ex10	Nessie :)

A08	dy04	fox
A08	dy06	dog with flat tail
A08	dy07	racoon with tail parallel
A08	dy12	dog/rat with straight tail
A08	es02	wee fox
A08	es03	the fox facing right
A08	es06	the fox
A08	es07	fox with three legs
A09	dy01	llama (llama facing right)
A09	es01	the chair one with the triangles facing the ground
A09	es02	the giraffe facing right
A09	es06	the giraffe
A09	es12	throne
A09	ex02	the banner with the bite out
A09	ex05	Fat giraffe looking over shoulder
A10	dy05	jet plane
A10	dy13	that random shape
A10	es07	one with square triangle on the right
A10	es09	House headed creature
A11	es11	weird cat thing
A11	ex01	the dog with the tail
A11	ex02	cat
A11	ex04	the cat
A11	ex06	monekycat
A11	ex09	the resting animal with the tail underneath
A12	dy02	fox
A12	dy03	fox facing right
A12	es04	dog/wolf figure with parallelogram tail pointing upwards
A12	es09	Fox, tail up
B01	dy12	bird flying left with the trapezium head
B01	es07	the bird flying left with wing higher than head trapezoid head
B01	ex02	bird with big head
B02	dy01	Eagle facing right
B02	dy12	the totem pole facing right
B02	es02	bird flying up
B02	es04	impaled eagle
B02	es08	sitting falcon facing right with wings out sitting on a triangle
B02	es09	the eagle facing right
B03	dy04	left facing wedge tail bird
B03	dy08	The bird flying upwards towards the left
B03	es06	the bird with its left wing facing it
B03	ex03	big bird looking down
B03	ex11	bird with the unfortunate arm wings
B04	dy04	crane
B04	dy05	bird facing either way

B04	es07	straight neck swan flying left
B04	ex04	fat bird
B04	ex10	left sitting bird!
B05	es06	the triangle headed bird in flight
B05	ex03	bird flying left
B06	dy04	perched bird
B06	dy06	the hunched eagle standing on triangle
B06	dy07	the vulture
B06	es04	hunchback not falling corw*crow
B06	es11	bird on perch. looking at its its left hand profile the bird has a triangle head
B06	ex11	perched bird of prey facing left with the short back
B07	es11	NW bird
B07	ex01	bird flying toward top left corner both facing top left corner
B07	ex02	bird beak up
B07	ex03	bird with beak upwards and wings pointing in
B07	ex08	bird flying to the top left corner
B07	ex10	left/diagonal bird
B08	dy01	crow facing right
B08	es01	hunched bird facing left right!!!! not left
B08	es03	hunched up bird, facing right vulture
B08	ex11	the perched bird of prey facing right with the longer back
B09	dy05	bird facing either way
B09	dy08	The fat bird going right.
B09	dy12	bird flying right
B09	es06	the swan facing right
B09	ex09	swan facing right
B10	dy02	german eagle
B10	dy03	american bird
B10	dy08	The bird facing left perched on a triangle with 2 triangle wings facing upwards
B10	es04	impaled eagle facing left
B10	es07	flying facing left flying up*
B10	es09	Eagle facing left
B10	es11	bird with wings outstretched mayan symbol one triangle perch, looking to the left with two triangles pointing up on either side for wings
B10	ex06	eagle standing straight on a upwards triangle left
B10	ex09	the bird the statue facing left
B10	ex10	Aztec bird
B11	dy01	left facing crow
B11	es01	hunched bird facing left hunched bird facing left
B11	es03	the vulture bird facing left
B11	es08	the birdseye view
B11	es12	big eagle

B12	dy02	Thin bird
B12	dy03	the bird flying to the left with a skinny cut arrow bodyhead is a triangle on the left side
B12	es01	bird flying left with triangle pointing down
B12	es12	duck horizontal
B12	ex05	small bird flying left
B12	ex06	small bird like the same bird, small bird one
B12	ex08	the little bird
P01	es02	the tissue man
P01	es04	spiritual figure holding lantern (triangle pointing towards himself)
P01	es08	praying with traingle hanging left
P01	es11	kneeling man with triangle
P01	ex02	kneeling man with triabngle dangling
P01	ex05	woman in long dress holding triangle
P01	ex09	the kneeliing man dropping the triangle
P02	dy01	Right facing angel oops! Left!
P02	dy05	praying man with crab claw
P02	ex05	rooster looking left
P02	ex10	I have kneeling man with pointy object
P03	dy06	man standing with cup
P03	dy07	priest offering
P03	dy13	the man serving the bowl facing left
P03	es04	Soup man
P03	es06	triangle-hat man holding triangle-vase
P03	es11	man holding soup
P03	es12	chef
P03	ex04	dude with bowl
P04	dy02	the guy with the missing hip
P04	dy03	dancing person with arms to the right
P04	es06	the guy on the two-legged stool
P04	ex02	second one is the person with a body like a road arrow
P04	ex06	the man stading straight with two legs and this in the arm <
P05	dy04	kneeling person with triangle pointing up
P05	dy06	man kneeling (triangle pointing up)
P05	dy07	running man
P05	es12	guy with laptop
P05	ex06	traingle towards running man
P05	ex08	the guy with the triangle level to his arm pointing right
P05	ex10	waiter carrying triangle tray
P05	ex11	man on one knee facing right and holiding out the triangle above his hand
P06	dy08	The one that looks the zombie with its arms outstretched. Big foot out and little foot behind/
P06	dy12	the man pointing to the left skinny body
P06	es01	standing man one foot infront of other

P06	ex01	inverted c man
P07	dy02	proposing man
P07	dy03	praying man
P07	dy05	pointy back foot praying man
P07	dy13	the man praying with the triangles on the bottom
P07	es06	praying man standing in a triangle basket
P07	es07	kneeling chunk
P07	ex04	the man sitting on triangles with his arms raised to the right
P08	dy05	normal praying man
P08	dy08	The man kneeling towards the right praying with diamond head
P08	es01	man kneeling right with long legs and looks like he's reading a book
P08	es07	praying no knee missing
P08	es12	praying guy with legs
P09	dy01	gown pleader facing left
P09	dy02	the one with facing left with one triangle like barely attached
P09	dy03	ghost with lump
P09	dy08	The zombie going left.
P09	es03	the praying man with the two triangles for arms facing left also
P09	es04	fat man with 2 triangles our of his back
P09	es07	big belly big belly
P09	es09	Angel facing right
P09	ex08	axe men
P09	ex11	that man again knees 2 arms left
P10	dy06	the kneeling one one leg with triangle pointing up
P10	dy07	running man with triangle below
P10	es02	the running guy
P10	es03	the running man facing right holding the triangle
P10	ex01	(detached triangle underneath arm) to be known as 'detached triangle man' :P
P11	dy13	man with hands facing downward, not holding triangle
P11	es09	man with small triangle left leg
P11	ex01	the NON detached triangle man with a big triangle for his right legi don't know what to call him... maybe bob or something...
P11	ex02	dog begging for food
P11	ex04	the praying mantis hand guy sitting on triangles his hands are kinda pointing down to the right diamond head, sitting on 2 triangles, the small one on the left right angle, big one upside down isoceles on the right
P11	ex05	dog runnnig up right wall
P11	ex09	cheeky wee scamp wee triangle for left foot, large for right
P11	ex10	man with a little pointy foot
P12	dy04	distorted stick figure
P12	es03	the two-armed running man
P12	es09	Guy with parallellogram left arm

P12	ex03	happy dog/horse man
P12	ex09	runny jumpy guy the runny jumpy lad with the fat right arm and little left
T01	dy02	the flame on the triangle pointing down with unattached the two triangles and rectangle in between underneath
T01	dy03	floating candle
T01	es02	nose and mouth
T01	es09	Oil lamp
T01	ex03	Floating dish
T01	ex08	the mohawk guy
T02	ex03	dish no gaps
T03	dy06	upside down man
T03	dy07	sting ray with big tail
T03	es01	downward bird with the triangle between the long legs
T03	ex01	the guy on his head
T03	ex04	fighter plane
T04	dy05	the three pointy bits with the diamond on top with two squares at angles
T04	dy06	the yay man
T04	dy07	funky diamond W
T04	ex01	the flame with the three peaks
T04	ex06	the w with the diamond off centre
T04	ex08	the fire pit with the diamond balancing on the three diamonds
T04	ex09	crown, with flame above
T05	dy13	upside down man with titled head
T05	es02	bull with attached square
T05	es08	batman with right cheek cut
T05	ex05	Devil face
T06	dy12	the fishy one
T06	es11	2 triangles. hex body. square on top
T06	ex10	squares with fins, flame, square ontop
T07	dy08	the zag zaggy thing that looks like nothing with the diamond
T07	es04	candle with parralelogram flame(outward triangles)
T07	es08	middle fimger
T07	es12	candle
T07	ex04	toblerone
T07	ex05	three mountains with candle above
T08	dy05	and the squares that meet in the middle
T08	dy13	that butterfly
T08	es03	a butterfly like with two ‘squares’, a parallelogram with nothing on top
T08	es07	flame with fish
T08	es11	2 hooked squares with diamond ting atop
T08	ex04	the two fish swimming away from each other

T08	ex09	arrows pointing outwards, flame in the middle like two outward facing arrows, with the parrellelogram flame bit
T09	dy02	shape with two triangles on either side and <=>in between
T09	dy03	the mushroom
T09	es09	skull
T09	ex02	the perfume bottle turned upside down
T09	ex03	apaceship spaceship
T09	ex11	the skulls
T10	es01	candle with the outward pointing triangles the candle with the 2 triangles
T10	es03	candles light with two triangles pointing outward big candles with big triangles candle*
T10	es08	candle
T10	ex02	monument
T10	ex05	Candle between two triangles pointed outwards
T10	ex06	the other bra
T11	dy04	upside down skull
T11	es04	candle without flame
T11	es06	the starward chimney ship *starwars
T11	es07	hexagon box on top
T11	es11	triangles to the side, HEXAGON middle, square top
T11	ex02	perfume bottle upwards
T11	ex11	upside down skull upside down skull the skull upside down
T12	dy12	the goblet one
T12	ex06	bra
T12	ex08	trophy

Table 11: Descriptions for naive raters. All are from the fourth time an image was directed. "dy" indicates a group in the Dyad condition, "es" the Esoteric triad, and "ex" the Exoteric triad.