THE UNIVERSITY *of* EDINBURGH

Language adapts to pressures from production: Experimental and computational evidence

Aislinn Keogh

Doctor of Philosophy

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

2025

Abstract

Languages have the daunting task of conveying an infinity of possible meanings. Yet at the same time, they are constrained by the limits of human memory, perception, and motor control. In this thesis, I study how language structure emerges from a complex interplay between these constraints. I argue that language is shaped by pressures for ease of retrieval and articulation, pressures which stem from the cognitive demands of real-time language production. Moreover, these pressures do not always pull in the same direction as those imposed during learning or comprehension. Using a combination of empirical methods, I show that, both at an individual-level and at a population-level, language structure is fundamentally a balancing act between competing pressures.

First, in Chapter 2, I investigate the mechanisms underlying *regularisation*, a well-documented process whereby languages become less variable over time. I test whether regularisation behaviour is driven by memory limitations during language learning or language production. In an artificial language learning experiment, I show that taxing working memory during production results in the loss of both predictable and unpredictable variation. Using a computational "urn" model, I demonstrate that a simple self-priming mechanism can generate this same pattern of results. However, I also find that the process by which random variation becomes more predictable is better explained by learning biases than by online production effects.

In Chapter 3, I adopt a more unified view of learning and production, by considering how language production might itself shape language learning. In collaboration with my colleague Elizabeth Pankratz, I test whether practising a new language with a more active, production-like task — compared to a more passive comprehension task — can help adults acquire a hidden morphological rule from underspecified input.

In two artificial language learning experiments, we find that participants clearly preferred not to segment below the word-level, regardless of task or prior experience with similar morphological rules.

Finally, in Chapter 4, I return to the question of how production pressures drive processes of language change, focussing on the role of communication in the evolution of lexicon structure. A naive view of communication might predict that words within a language would be as different from each other as possible to avoid potential confusion. However, compared to a range of random and phonotactically-controlled baselines, I show that words are actually more *similar* to each other than would be expected, a property I refer to as *phonetic clustering*. In an agent-based exemplar model, I show that this property arises from a trade-off between opposing forces: production biases select for increased similarity between words, whilst a comprehension mechanism works to maintain distinctiveness. I then simulate these same pressures in a series of communication experiments to investigate how language users adapt their lexical choices to facilitate efficient communication. Again, I show that they strike a balance between ease-of-production and ease-of-comprehension, although they do not always converge on the most optimal solution.

Overall, this thesis sheds light on how pressures from language production compete and cooperate with other selective pressures to shape language learning and, ultimately, language structure.

Lay summary

Humans are incessant chatterboxes. Whether we're speaking, signing or writing, complex communication is one of the most quintessentially human behaviours. And since we spend so much of our lives talking, it's tempting to think that it's a pretty trivial task — something we do effortlessly.

In this thesis, I argue that this task — language production — is actually far from effortless. Rather, it comes with a whole host of challenges: dredging words out of our long-term memory, arranging them into the right order in our heads, and finally getting them out of our heads to be pronounced by our mouths or hands. This is a highly intricate sequence of events, and things can go wrong at any stage. The central idea of this thesis is that languages look the way they do because they *adapt* to these difficulties.

In three projects, I study how the challenges associated with language production shape the way we learn languages, the way we use them in real-time communication, and the way they change and evolve over time. I do this in two main ways. The first involves asking human participants to learn miniature made-up languages, and then seeing what happens when they have to produce these languages themselves. The second involves using computer code to *simulate* the process of human communication, to see how the effects I observe in my experiments might accumulate over thousands of generations.

Overall, through this two-pronged approach, I offer converging evidence for my core proposal: that languages evolve to become easier to produce, all the while maintaining their incredible ability to convey an infinity of possible meanings.

Acknowledgements

It's alright, children. Life is made up of meetings
and partings — that is the way of it.

Kermit the Frog, The Muppet Christmas Carol

As is often the case with these things, it was really just a series of serendipitous events that put me on the path to this PhD. The first was a lecture about vervet monkeys in Newcastle, after which I begged Maggie Tallerman to let me switch supervisors to do my undergraduate dissertation with her. The second was a chance meeting with Joel Wallenberg in the entrance hall of the Percy Building, where for no obvious reason, I asked him if he wanted to hear about said undergraduate dissertation (and then he told me about information theory). And the third, several years later, was an article about fricatives and farming in the New Scientist, which prompted an email to Seán Roberts asking if we could meet up for a chat over a cup of hot leaf liquid. Thank you, all, for saying yes to my brazen demands on your time.

I've been profoundly fortunate to be supervised by two extremely smart and extremely kind people: Simon Kirby and Jenny Culbertson. Since I first met Simon in 2019 when I was just vaguely thinking about applying to the Masters, he has had a seemingly unshakable confidence in my ability to succeed (that I have not always shared). He has an endless supply of big ideas, but he's always the first to tell you how cool yours are. Jenny is ridiculously on top of things, and is someone you can truly depend on through the good times and bad. She has an incredible eye for the finer details, and her feedback has immeasurably improved me as a researcher. Thank you, both, for everything.

To my PhD twin, Elizabeth, thank you for being a constant presence on this journey.

We've come a long way from our first year lying on the floor of the 40 George Square offices trying to work out what we even wanted to do with our lives. I can't imagine what the PhD would have looked like without you, and I'm glad I never had to find out. And for anyone who enjoys the cute illustrations in this thesis, you should know that they're almost all Elizabeth's handiwork.

I truly landed on my feet in joining the Centre for Language Evolution. From the first picnic in Holyrood Park, I knew that these were my people[1]. Special thanks to the first friends I made in Edinburgh, Maisy and Annie, for riding the rollercoaster with me this whole time, and for making my life infinitely more joyful, silly, and full of lil guys. Shout out to the group supervision crew: Elizabeth, Annie, Marc, Fang and Sabine. There were some tears (often mine), but we made it through together! Thanks to my fellow students (CLE and otherwise) who have made my time in the chatty office all the more chatty, including but not limited to Maisy, Elizabeth, Lucie, Sabine, Ari, Vilde, Elif, Richard, Gilly, Fang and Federico. And to all the other CLE and CLE-adjacent folks who have been there along the way — Juan, Shira, Jo, Ponrawee, Matt, András, Henry, Lauren, Yevgen, Patrick, Anna, Richard — thank you for making this such a wonderful community; you're all dope.

Before September 2020, I had never seen a programming language. A number of people have been instrumental in getting me to see several without imploding. Alasdair Tullo is paid to help people with this stuff, but has been far more enthusiastic about doing so than that position really required (including when I brought him problems that were combinatorically impossible). Kenny Smith is paid to help *some* people with this stuff; I'm not one of those people, but he very kindly never mentioned this when I asked him how to achieve a new kind of jsPsych wizardry. Sam, Dan and Rich are *definitely* not paid to provide my tech support, but have done so unquestioningly.

I have been supported in more abstract ways by many friends — academic and real-life — but two very special friends deserve a special mention. Hannah has wanted me to do a PhD since approximately the second day of our friendship, and I'm delighted to have finally made her dream come true. And Rich, despite living over 300 miles

[1]This feeling has been reinforced many times in the last five years, not least when I was presented with cake in the form of a probabilistic urn model after my viva.

away for most of my PhD, has been at every one of my CLE talks, and despite being a software engineer and not a linguist, has always asked (annoyingly) smart and insightful questions that have made a genuine contribution to my research. Thank you, both, for your unwavering love and support.

Sadly, two other very important people passed away before they could see me complete this thesis, so I'd like to raise a metaphorical glass to them. To my godfather, Tony, who I have no doubt would have loved another opportunity to make an on-stage joke about my educational achievements. And to my grandmother, Anne, who was the epitome of a strong independent woman, and who instilled in me exactly the kind of self-determination I needed to get through a PhD.

A wise man[2] once told me that it's good for academics to have a hobby they enjoy but are not very good at. To that end, I have made damn sure to never garner any appreciable level of skill in skateboarding or climbing, despite spending many hours doing both over these last 5 years. And it turned out to be very wise advice indeed, since the only times I was really able to switch off my research-brain was when I was rolling around on a plank of wood or halfway up a wall. So a huge thank you to all my skating and climbing pals in Scotland, England and Wisconsin[3] for quite literally keeping me sane throughout this process.

Thank you to The Muppet Christmas Carol, for providing a treasure trove of incomprehensible references with which to fill my CLE talks (and now, this thesis).

And finally, to my husband, Sam. I've left the most important person for last because, despite my clear propensity for verbosity, I'm not sure how I'm going to find the words to adequately convey his contribution to this thesis or to my life in general. What I can say, with no exaggeration, is that I couldn't and wouldn't have done any of this without his support — emotional, financial, and statistical. Sam, I love you, and I hope that Half Man Half Biscuit write a song about how cool you are.

[2]Incidentally, the same wise man who once told me to stop discovering things (see Chapter 5).
[3]For the avoidance of doubt, I am aware that Wisconsin is not one of the member states of the UK.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It's all just production isn't it?!

Professor Jennifer Culbertson

Humans are constantly chatting. Whether we're speaking, signing or writing, complex communication is a hallmark of the human species. And since we spend so much of our lives talking, it can be tempting to think that it is a trivial task. Yet it takes years to progress from babbling to putting full sentences together, and even when we achieve "mastery" of a language, we continue to make mistakes: slips of the tongue, losing our train of thought, even using the wrong words altogether.

This thesis has at its core three key ideas: (1) that producing language in real-time comes with a host of cognitive and motor challenges, (2) that the challenges posed by production have significant implications for language structure, and (3) that language is ultimately shaped by the interplay between production pressures and other functional pressures arising from learning and comprehension. I will return to each of these ideas soon. But first, I want to take a moment to situate my research in its wider context.

1.1 A usage-based, cultural evolutionary approach

Fundamentally, in adopting the position I've just outlined, I am putting myself firmly in a camp which sees language structure as *emergent* from learning and use, rather than

as a primitive which is specified by domain-specific constraints (cf. Chomsky 1965). The view of language I take in this thesis is well expressed by Bybee (2010: 1):

> The structural phenomena we observe in the grammar of natural languages can be derived from domain-general cognitive processes as they operate in multiple instances of language use.

This is typically described as a *usage-based* perspective, one which assumes a profound relation between linguistic structure and usage (Coussé & Mengden 2014). Usage-based and other functionalist approaches have tackled the notion of language "structure" at a number of levels. At an individual level, people develop their own internal representation of the structure of their language(s) as they learn; in the usage-based tradition, this structure is seen as emergent from the acquisition process (e.g. Goldberg & Casenhiser 2008; MacWhinney 1998; Tomasello 2003), rather than depending on innate representations specified by a Language Acquisition Device (Chomsky 1965; Crain et al. 2017; Snyder 2007; Yang et al. 2017). At a population level, many aspects of these individual representations are conventionalised as the language's grammar; usage-based linguists are concerned with documenting these grammars not as a set of abstract rules, but rather in terms of the actual frequency with which particular constructions are used (e.g. Bybee 2007; Croft 2001; Goldberg 2005; Haspelmath 2008; Hopper 1987). Finally, at a species level, some structural features reoccur systematically across the world's languages; a usage-based approach sees these commonalities as emergent from general principles of human cognition, sociality, and communicative efficiency (e.g. Bybee 2010; Christiansen & Chater 2008, 2016b; Givón 1979; Hawkins 2004; MacDonald 2013; Schmid 2020; Zipf 1949).

My primary interest lies in this final piece of the puzzle: the overarching, big-picture properties that characterise human language as a whole. Kirby (1999) points out that there is a "problem of linkage" in the functionalist view of such features: it is true there is a striking fit between language's design and its function, but demonstrating the existence of such a fit does not, by itself, explain anything. But a large and growing research tradition highlights that an explanatory mechanism does exist, a mechanism by which processes of language use can give rise to language universals: *cultural evolution* (e.g. Arnon & Kirby 2024; Beckner et al. 2009; Boyd & Richerson 1988; Cavalli-Sforza & Feldman 1981; Chater & Christiansen 2010; Croft 2000; Griffiths et al.

2008; Hurford 1999; Kirby 2017; Kirby et al. 2007, 2008, 2015; Roberts & Fedzechkina 2018; Saldana et al. 2019; K. Smith 2011; K. Smith et al. 2003b, 2003a; Spike 2016; Steels 2011; Thompson et al. 2016). On this view, language evolution is essentially a process of selection between competing variants: as languages are passed from person to person, variants that are more easily learned, processed or produced will tend to win out. In Chapters 2 and 4, I pursue this cultural evolutionary approach, with a particular focus on how pressures imposed by language *production* shape language form (MacDonald 2013).

Chapter 3 represents something of a departure from the evolutionary perspective of the other content chapters, instead examining how individuals learn the rules of a new language. At first blush, this is a very different question than the question of how those rules got there in the first place. However, Chater and Christiansen (2010) point out that it may not be necessary to postulate a sharp distinction between acquisition and evolution. Instead, they argue for an integrated framework (Christiansen & Chater 2016a) which sees acquisition, processing *and* evolution as closely intertwined, treating them simply as different timescales on which humans create language. My interest in this middle chapter is in how language structure within individuals' minds might emerge under the same pressures and processes that shape language at a population or species-level. Again, my primary focus is on the role of language production.

In what follows, I set out my motivation for adopting this focus on language production, outlining in more detail the three key ideas that have guided the work in this thesis. Then, in Section 1.3, I summarise the methodological framework I have used to investigate the emergence of language structure on different timescales. Finally, in Section 1.4, I provide a roadmap for the rest of the thesis.

1.2 Three key ideas

1.2.1 Language production is hard

Producing language in real-time requires us to rapidly turn our mental representations into meaningful output. And although we may seem to do this almost effortlessly, it

is actually a highly complex behaviour. To produce an utterance, we need to retrieve the right units to convey our intended meaning, assemble these units into an ordered sequence, and then send this sequence to our motor articulators to be pronounced (Bock 1995; Levelt 1989). To make matters even more complex, all of these processes are likely happening concurrently: while we are producing one utterance, we are already planning the next one (F. Ferreira & Swets 2002). And crucially, the whole production pipeline — retrieval, planning, articulation — is constrained by the limits of human memory, attention, and motor control. Put simply, we need to finish producing an utterance before we forget (or lose interest in) what we were trying to say. And we can only achieve that goal according to how quickly and skillfully we can manipulate our articulatory apparatus (for speech, see e.g. Fitch 2010: Chapters 8-10; A. Smith et al. 1995; for sign, see e.g. M. J. L. Gómez et al. 2007; Poizner et al. 1983).

There are a huge number of ways in which production can go wrong. When trying to retrieve a lexical item from long-term memory, we sometimes activate a slightly different one than the one we were aiming for (Dell 1986; Goldberg & Ferreira 2022; Koranda et al. 2018; Levelt 1999; Roelofs 1992). Sometimes, we can't retrieve a particular lexical item at all, even if we have access to some of its features — a "tip-of-the-tongue" state (A. S. Brown 2012; Cleary 2017; Schwartz 2002; Vigliocco et al. 1997, 1999). When it comes to assembling lexical items into utterances, planning burdens have the potential to give rise to syntactic errors or word order variations that alter the intended meaning (Deese 1984). We are also highly influenced by the words, phrases and abstract structures we have heard or uttered recently, and prone to re-using these even if they might not be the most appropriate for the current context (Koranda et al. 2020; Lee et al. 2022). And finally, once we have an utterance plan ready to produce, it is vulnerable to mishaps in articulation — slips of the tongue, like sound exchange or anticipation errors (Dell 1986; Shattuck-Hufnagel & Klatt 1979; Stemberger 1990). Although outright errors are relatively rare (Bock & Levelt 1994; Heeschen 1993), speech is certainly punctuated by disfluencies: fillers like *uh* and *um*, word prolongations, repetitions and pauses are all common (Corley & Stewart 2008; Engelhardt et al. 2010; V. S. Ferreira & Dell 2000; Finlayson & Corley 2012; Fox Tree & Clark 1997).

All this is to say that production is not trivial: language users face a myriad of difficulties in getting from a thought to a successfully transmitted utterance.

1.2.2 Production difficulty has implications for language structure

Despite the vast literature documenting the challenges associated with production, explicitly production-based accounts of language structure are hard to come by (although see e.g. Bock 1982; Ohala 1993). However, one theory that has been very influential on my thinking is the Production Distribution Comprehension (PDC) account (MacDonald 2013). The PDC contends that "language producibility, more than learnability or comprehensibility, drives language form" (2013: 13). The crux of the argument is that the way we say things is shaped primarily by the (subconscious) choices we make during production to make our own lives easier, rather than by our efforts to make things easier for our conversational partners. My own view is that producers clearly have communicative goals which involve being understood, but to the extent that we can ease the burden of production whilst still achieving these goals, we will take opportunities to do so.

So how might our attempts to ameliorate production difficulty affect what actually gets produced? The PDC's focus is on the difficulties of utterance planning, and how biases that promote greater production fluency might shape patterns of sentence structure. For example, a bias to produce more easily retrieved words earlier in the utterance — buying time to retrieve less accessible items — can give rise to word order variation, such as the choice between active and passive voice (Bock 1982, 1995; Bock & Warren 1985; V. S. Ferreira 2008; Tanaka et al. 2011). Conversely, a bias to re-use recently or frequently executed utterance plans (i.e. priming, also known as structural persistence) can lead to greater *rigidity* in word order, even in languages that license relatively free word order (Bock et al. 1986; Christianson & Ferreira 2005; V. S. Ferreira & Bock 2006; Mahowald et al. 2016). Although not within the scope of the PDC, the challenges associated with motor articulation also have clear consequences for language form. For example, producers tend to try and minimise articulatory effort by pronouncing words less carefully or truncating them (also known as *clipping*), especially words that are more predictable and thus convey less information (Aylett & Turk 2004; Bybee 2002; Hall et al. 2018; Jamet 2009; Kanwal et al. 2017; Mahowald et al. 2013; Pierrehumbert 2001; Stamp et al. 2024; Zipf 1949).

For me, these observations have implications for our understanding of language

structure on multiple timescales. In terms of individual learning, if we're repeatedly making the same kinds of choices in production, this is likely to feed into our mental representation of our language(s). Specifically, things we produce more often will be strengthened in memory, thus becoming ever more likely to be used again in future productions — a self-perpetuating process (Karpicke 2012). And on an evolutionary timescale, the effects of individual-level behaviour can accumulate (Kirby et al. 2007; K. Smith 2011). This latter point is at the heart of the Distribution component of the PDC: "Summed over millions of utterances and many language producers, implicit production choices favoring less-difficult forms create dramatic statistical regularities in language usage" (MacDonald 2013: 5).

A famous example of such a statistical regularity is the Law of Abbreviation, a cross-linguistic tendency for more frequent words to be shorter (Zipf 1949). Zipf argued that this property arises under a Principle of Least Effort, whereby producers shorten words wherever possible to minimise articulatory effort. Other core properties of language may also plausibly trace their origins to production processes. For example, function words and grammatical markers often develop from older lexical items through *grammaticalisation*; priming has been proposed as an underlying mechanism in this process (Jäger & Rosenbach 2008). Ambiguity — which is a pervasive feature of language — makes production more efficient by allowing for the reuse of more easily articulated words and sounds (Piantadosi et al. 2012). And there is some evidence that phoneme inventories tend to be organised around more easily-articulated consonants in speech (Everett 2018) and handshapes in sign (Ann 1996).

In sum, I think there is good reason to adopt a stronger focus on production mechanisms when developing theories of language acquisition and evolution; after all, the only way we generate the linguistic data that feeds these processes is through what we produce.

1.2.3 Language is shaped by competing pressures

However, the idea that production drives language structure is not the whole story: clearly, languages do not evolve *only* to be easily producible. Rather, they are subject to a vast array of cognitive and social pressures. They must allow us to convey an infinity

of possible thoughts in a way that can be decoded by the people we're talking to. They must be learnable — by infants and, in some cases, by adults. They must allow us to signal social group membership, yet also talk to people with whom we share little common ground. They must be responsive to changes in culture and society that create new words or displace existing words. These various pressures sometimes work in harmony, and sometimes pull against each other. As Beckner et al. (2009: 2) put it, "a speaker's behavior is the consequence of **competing factors** ranging from perceptual constraints to social motivations" (emphasis added).

To illustrate this point, let us return to the Law of Abbreviation, which states that more frequent words tend to be shorter (Zipf 1949). If ease of production was the only factor shaping languages, we might expect *all* words to be as short as possible: all else being equal, shorter words take less effort to produce. However, shorter words are more likely to be lost in noisy transmission, or to be outright ambiguous (since there are fewer possible unique short words); there is therefore a countervailing pressure from comprehension in favour of long words. The Law of Abbreviation offers a compromise between these competing pressures: it means that the words we say a lot are optimised for production ease, while the words we say less often (which are also likely to be more difficult to process: Brysbaert et al. 2018) are optimised for successful comprehension. In other words, words are shorter where possible, and longer where necessary.

Similar trade-offs have been invoked to describe many core features of language and recurrent pathways of language change, including compositionality (Beckner et al. 2017; Kirby et al. 2008, 2015; Saldana et al. 2019; K. Smith et al. 2003a), patterns of word and morpheme ordering (Christensen et al. 2016; Gibson et al. 2013b; Hahn & Yang 2022; Hahn et al. 2021, 2022; Holtz et al. 2023), the organisation of semantic category systems (Hallam et al. 2025; Kemp & Regier 2012; Regier et al. 2015; Xu et al. 2020) and sound category systems (Wedel 2012; Winter & Wedel 2016), homophony and other kinds of ambiguity (V. S. Ferreira 2008; Piantadosi et al. 2012; Trott & Bergen 2022), semantic extension (Harmon & Kapatsinski 2017), phonetic reduction (Aylett & Turk 2004; Bell et al. 2009; Hall et al. 2018; Lindblom 1990), sound symbolism and iconicity (Dingemanse et al. 2015; Jee et al. 2022; Monaghan et al. 2014), and various kinds of grammatical marking (Fedzechkina & Roberts 2020; Roberts & Fedzechkina 2018; Seržant & Moroz 2022).

The take-home message here is that competition between different usage pressures is inevitable: as complex adaptive systems (Beckner et al. 2009; Bybee 2010), languages must find ways to strike a balance between pressures that pull with different strengths in different directions. My main focus in this thesis is on the role of production pressures in shaping language structure, a topic I believe has been under-explored thus far. Crucially though, I aim to understand how these pressures compete, cooperate, and coexist with those arising from acquisition and comprehension. To my mind, it is ultimately through this tug-of-war that language structure emerges.

1.3 Methodological framework

Having established that my goal in this thesis is to study the emergence of language structure, the obvious question is: *how*? Historically, usage-based linguists have relied heavily on cross-linguistic and historical corpus data. The idea is that features which appear in the synchronic record for a diverse sample of languages — or arise repeatedly through diachrony — may reflect cognitive, perceptual or pragmatic processes that are shared across human populations. Sometimes, corpus data is supplemented by evidence from psycholinguistic experiments showing that a particular feature which is shared across many of the world's languages confers a processing advantage compared to alternative structures. However, this juxtaposition does not, by itself, provide *causal* evidence for a relationship between human cognition and language structure: the problem of linkage remains (Kirby 1999).

To illustrate concretely, consider the case of dependency lengths. Dependency length is defined as the total linear distance between all syntactic heads and their dependents in a sentence; some examples are given in Figure 1.1. The dependency length minimisation (DLM) hypothesis states that language users prefer word orders that minimise dependency length, and that this should be reflected in grammars that facilitate the production of short dependencies through their word order rules (Gibson 1998, 2000; Hawkins 2004). We know that dependency lengths are generally shorter in real languages than would be expected if there was no pressure for DLM (Futrell et al. 2015; Gildea & Temperley 2010; Hawkins 2004; Liu 2008; Temperley 2007; Yadav et al.

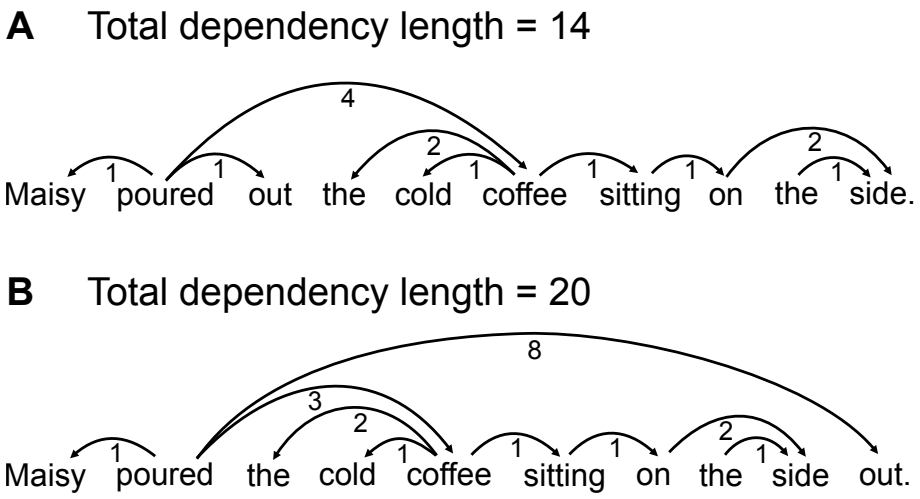

Figure 1.1: Dependency representations for two sentences with the same semantic interpretation, but different word orders. The number under each arc represents the length of that individual dependency; the total dependency length for each sentence is given by summing all these numbers. English speakers typically find a sentence like A — which has a shorter total dependency length — more natural than one like B (Futrell et al. 2015).

2022). There is also good evidence that long-distance dependencies are a source of processing difficulties, both in production and comprehension (e.g. Fedorenko et al. 2013; Gordon et al. 2002; Grodner & Gibson 2005; McElree et al. 2003; Momma 2021; Nicenboim et al. 2015; Traxler & Pickering 1996; Van Dyke & Lewis 2003; Yamashita & Chang 2001). However, these two facts alone cannot conclusively demonstrate that language structure is shaped by human information processing, and indeed, the causality could even go in the other direction. In other words, it could simply be that people struggle with long-distance dependencies because such structures are relatively uncommon, so there are limited opportunities to learn to process them. To support a directional hypothesis, we would need evidence that directly links behaviour in individual language users with emergent structural features; in the case of DLM, this evidence is provided by studies which show that natural-language-like dependency lengths emerge through learning and use of artificial languages which do not initially conform to the hypothesised bias (e.g. Davis & Smith 2023; Fedzechkina et al. 2018).

This example demonstrates the core of the methodological framework I aim to pursue in this thesis: identify a linguistic phenomenon of interest, establish the processing mechanisms that underpin it, then probe the existence of a causal and directional relationship between these two elements. The benefits of such a multi-faceted approach

are underscored by MacDonald (2013: 1-2), who highlights a potential obstacle to developing usage-based theories of language structure:

> Language researchers must develop an account of the effects of experience on perception, but [...] must also consider why the experience — the language — has the character it does. This difficult task is compounded by the fact that the psycholinguists who study language use are typically not the same people as the linguists who study the nature of language form, so there is a gulf between linguistic theories of the nature of language and psycholinguists' accounts of the effects of experience with language patterns.

If I were to summarise my overall ambition in this thesis, it is that I have tried to be exactly such a researcher who does both: a psycholinguist who studies language use, and a linguist who studies the nature of language form. To provide the missing link between behaviour and typology, I use two core methods: artificial language learning experiments, and agent-based computational models. In what follows, I provide a brief summary and evaluation of these methods and the relationship, as I see it, between them.

1.3.1 Artificial language learning experiments

Natural language data — from acquisition, use, and typology — is an important source of evidence in the study of language structure. However, relying solely on this kind of data has clear limitations. To give just a very quick rundown of the key issues (discussed at greater length in Culbertson 2023 and Fedzechkina et al. 2016):

1. It is impossible to get a complete picture of people's experience with particular features, so we cannot tell which aspects of their language usage are driven by underlying cognitive, perceptual or motor constraints, and which are simply driven by the statistics of the linguistic input they have been exposed to.

2. Languages are complex organisms with many moving and interacting parts, so it is inconceivable that two languages will ever differ *only* in terms of the particular feature we are trying to study. Rather, they are likely to vary along many dimensions, including their social and pragmatic context.

3. Languages are deeply inter-related — both phylogenetically and areally — so some commonalities between them may simply be accidents of history. These

relationships also dramatically reduce the effective sample size, since related languages cannot be considered as independent data points.

4. As discussed earlier, even if some features appear to be very frequent and robust cross-linguistically, this tells us nothing about their *origin*.

Taking language into the laboratory offers a more controlled environment in which to test specific hypotheses about the relationship between human cognition and language structure. Behavioural experiments using miniature artificial languages — designed to isolate the phenomenon of interest and minimise the effect of other language experience — have a long and rich history in developmental psychology (e.g. Braine et al. 1990a; R. L. Gómez & Gerken 2000; Reber 1967; Romberg & Saffran 2010) and have, more recently, been adapted to investigate what kinds of cognitive constraints shape language structure (for reviews, see Culbertson 2023; Fedzechkina et al. 2016). This method allows researchers to design linguistic systems with particular properties, and to exercise control over participants' experience with these properties.

Typically, artificial language learning experiments are used — unsurprisingly, given the name — to ask questions about *learning*: for example, are typologically more frequent patterns easier to learn? However, the terminology is arguably slightly misleading: different paradigms under this umbrella differ both in the extent to which learning is actively involved, and in the extent to which they really isolate learning as the explanatory mechanism (Culbertson 2023). For example, a hypothesised bias concerning the relative order of nominal modifiers (Universal 20: Greenberg 1963) has been tested in a series of artificial language learning experiments (Culbertson & Adger 2014; A. Martin et al. 2020, 2024). The results clearly support the existence of an underlying cognitive bias in this domain: across different paradigms and different native speaker populations, participants strongly preferred the typologically more common orders. Crucially though, the effect was strongest when participants had to *produce* noun phrases themselves.

To me, this suggests that "learning" is not the whole story here: it is not just that participants *learn* certain orders more easily than others, but also that they *produce* certain orders more readily than others. Of course, production and learning are not two completely different processes: what is produced is, at least partially, a reflection of

what has been learned. Nonetheless, it may be the case that certain structural alternations can be equally well learned and stored in memory, but that one confers an advantage specifically during production — perhaps because of constraints on utterance planning, or because of accessibility effects in memory retrieval (Goldberg & Ferreira 2022; MacDonald 2013). Therefore, in my experiments, I aim to tease apart the different mechanisms at play, and specifically, to pinpoint the effects of online production *after* taking learning effects into account.

Finally, a note on modality. Across the three projects presented in this thesis, my experiments are designed as proxies for spoken language[1]: I present the artificial languages primarily in writing, and have participants "produce" the language themselves by clicking buttons to assemble utterances out of smaller pieces (words, syllables, or letters). My reasons for taking this approach — compared to using auditory stimuli and oral production — are chiefly practical, but also philosophical.

On the practical side, almost all of the experimental data I present in this thesis has been collected at a distance: participants access the experiments from their own home, using their own devices. The benefit of this approach is that it has allowed me to collect relatively large sample sizes very quickly. However, it is also impossible to control the quality of participants' audio equipment (or indeed, even to ensure that participants comply with instructions to *use* audio equipment in the first place). With written stimuli, we can be more sure that different participants have access to the same information. Furthermore, the data that I collect from button-clicking tasks requires no or minimal manual coding, and is therefore quicker and easier to analyse than any kind of free production data — especially recorded speech.

The other side of equation is my general philosophical stance on the purpose of experiments. I want to be able to isolate specific mechanisms that drive the effects I'm interested in, rather than just pointing to "production" in general. To do this, my experiments are designed to *simulate* the pressures imposed by particular production mechanisms, rather than allowing these pressures to emerge organically (see e.g. Kanwal 2018 for other examples of experiments in this style). With oral production, it is far

[1]It's worth saying that, in general, I would expect production pressures to have similar implications for signed languages. Of course, there are also likely to be some factors that are unique to particular modalities; this would certainly be a fruitful avenue for future research.

more difficult to turn particular mechanisms on and off in the same way, or to pinpoint the effects of one mechanism over another.

Of course, I do acknowledge that the decision to use written stimuli and button-clicking tasks is a simplifying one, and therefore comes with its own limitations. I discuss the potential consequences of this decision throughout the three content chapters.

1.3.2 Agent-based computational modelling

Alongside artificial language learning experiments, Chapters 2 and 4 also include agent-based computational models. My aim with these models is to implement production pressures in a way that models oral production more closely (and in greater detail) than the button-clicking tasks in my experiments. The models I use are rooted in the *exemplar* framework (Bybee 2010; Nosofsky 1988; Pierrehumbert 2001; Shi et al. 2010; Wedel 2006). Exemplar models assume that people have a rich, episodic memory for the perceptual details of their linguistic input. That is, an agent's internal representation of their language is made up of concrete exemplars of linguistic behaviour they have observed, not of abstract generalisations about that behaviour. This approach stands in contrast to, for example, a Bayesian model, where agents use the data they receive — along with their prior biases — to derive a probability distribution over abstract hypotheses about the language's structure. The distinction between the two types of models is exemplified in Figure 1.2.

Bayesian models are undeniably popular in the field of language evolution (e.g. Culbertson et al. 2013; Griffiths & Kalish 2007; Griffiths et al. 2008; Josserand et al. 2021; Kalish et al. 2007; Kirby et al. 2007, 2015; Navarro et al. 2018; Reali & Griffiths 2009; K. Smith 2020; K. Smith et al. 2017; Thompson et al. 2016), so my choice of exemplar models is somewhat unusual (although see Wedel 2006). There are several reasons I have favoured this style of model in this thesis. First, I think they are a particularly convenient and transparent way to model production mechanisms like memory retrieval and articulation. Second, it's very easy to model memory *limitations* and forgetting in an exemplar model, by simply deleting some of the data — either randomly or in a more targeted way. When agents only store an abstraction about the data and not the

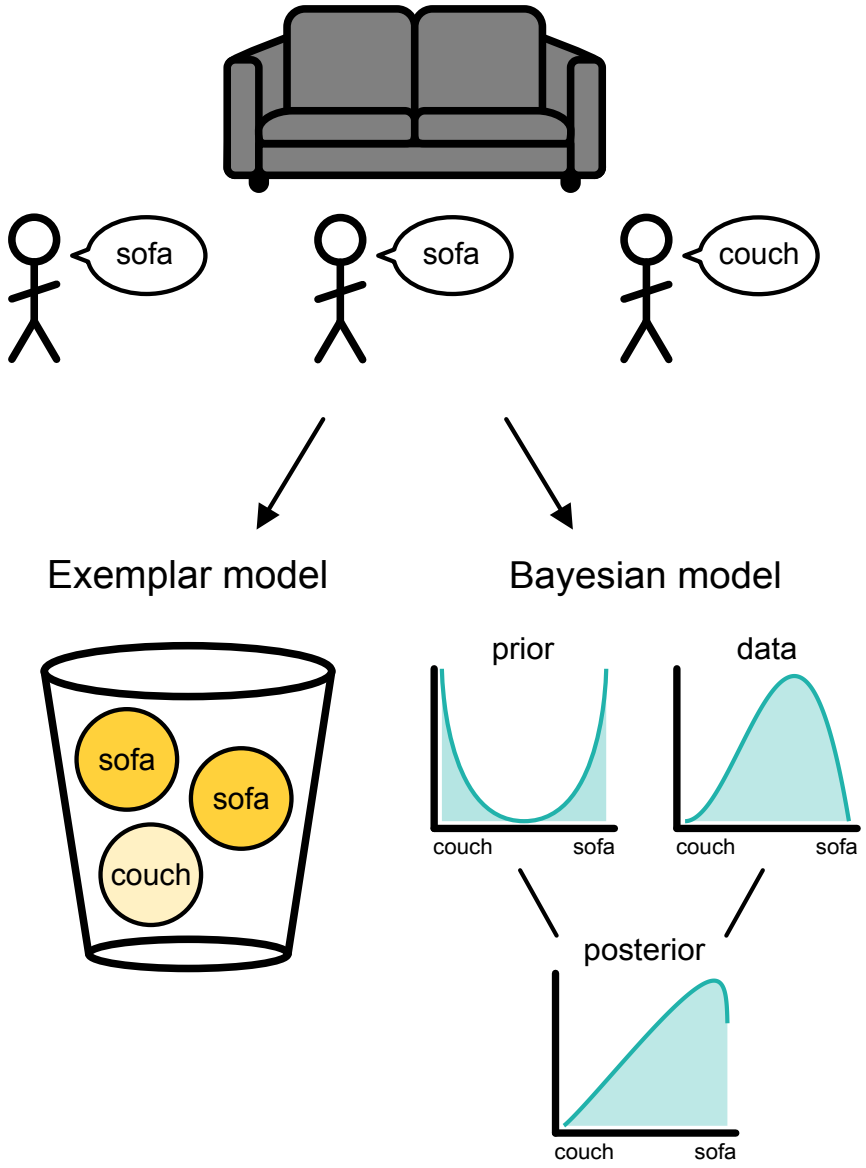

Figure 1.2: Example of the difference between an exemplar-based approach and a Bayesian approach to one simple inductive problem: learning about synonymy. The data that feeds in to both models is the same: two instances of the word "sofa", and one instance of the word "couch". In the exemplar model (left), the agent simply stores this data; there are no other layers of abstraction. To produce a word themselves, they would then sample (with replacement) from the stored data; in effect, this means they would produce "sofa" with probability 0.67, but this probability is never actually calculated. In the Bayesian model (right), the agent starts with some prior expectations about how object labels should be distributed: here, I am showing a u-shaped prior, which encodes the expectation that objects should only have one label (but it could equally well be either "sofa" or "couch"). Note that the x-axis is labelled with these words for ease of presentation, but in reality, this is a prior over θ: the probability of using some abstract word w. Alongside their prior, the agent derives a probability distribution over possible values of θ given the data they have received (the *likelihood* term in Bayes' Theorem). They then combine these two elements to yield a posterior probability distribution. To produce a word themselves, they would sample a value of θ from the posterior and produce the word "sofa" with probability equal to the sampled value (and "couch" with probability $1 - \theta$); the influence of the anti-synonymy prior means that they're likely to produce "sofa" slightly more often than they observed it.

data itself, as in Bayesian models, it is more difficult to imagine what memory decay would look like. And finally, exemplar models include all the ingredients we need for an evolutionary account: variation, reproduction, and selection (Wedel 2006; Winter 2014).

For me, the main appeal of Bayesian models is that they are a powerful way to study the relationship between *learning* biases (encoded in the prior) and emergent language structure. However, in this thesis, I am less interested in how learning mechanisms shape the representation we initially acquire, and more interested in how production mechanisms continue to shape this representation throughout our lives. Exemplar models are an ideal way to tackle this kind of question.

1.3.3 Why run experiments? Why build models?

I want to wrap up this methodology section by briefly considering the relationship between the two methods I've just described — experiments and models.

For me, the main benefit of experiments is that they get us somewhat closer to real language use than computational models, in that they use human participants. If we want to understand human cognition, it goes without saying that we need to study humans. This is important not least because human behaviour is noisy and idiosyncratic in ways that we might not think to build into our models, or might not even understand. For example, different people can be more or less motivated by different aspects of the task, more or less exploratory and curious, more or less sensitive to what they believe the researcher wants from them. Models have the potential to miss a lot of this inter-individual variation.

So why build models at all? One counterpoint to the argument I've just made is that the *way* people's behaviour is noisy in experimental settings might not be the same as in the real world. In fact, we only have to look to the last example I gave above to see that this is probably the case: sensitivity to a researcher's imagined desires is patently not a factor that shapes real language use (although sensitivity to an interlocutor's imagined desires might be). So perhaps, models allow us to gloss over some sources of inter-individual variation that are not likely to have strong explanatory power in the

real world. But probably the main reason to build models is a practical one: simply put, we can do things with models that would be infeasible with human participants. For example, models allow us to simulate many thousands of interactions or generations of language transmission, letting us observe effects that unfold over a much longer timescale than we could reasonably test in the lab. For me, the real beauty of computer models is that they also allow for a comprehensive exploration of cognitive architectures and mechanisms: we can use them to test out many small tweaks, fine-tuning our parameters until the model spits out something that looks like a real language. Critics might say that this is exactly why models are *not* informative: because we just "bake in" the result we want. But I would argue that this is missing the point. When we work with humans, try as we might to isolate the specific mechanisms we're interested in, there are always going to be a myriad of other factors that influence their behaviour. When we work with computer agents, we have full control over these factors, and can turn them on and off at will. This allows us to test precisely which mechanisms (and interactions between them) we need to include to observe the same kind of patterns as we see in our human data. In other words, an interesting "result" of a model might not the data it generates *per se*, but rather the unique combination of parameter settings which allow it to generate the right *kind* of data.

As I see it, experiments and computational models both have their place. Both are microcosms of the systems we really want to study, and as such, can provide converging evidence of the pressures that shape these systems. The benefit of combining these methods is that it allows us to demonstrate the existence of causal relationships between individual-level behaviour and emergent language structure, *and* to provide a mechanistic account of these relationships.

1.4 Thesis roadmap

This thesis comprises three projects, each probing a different aspect of linguistic "structure" (broadly construed). Across these projects, I use the methodologies outlined above to test hypotheses about the role of language production in the emergence of the phenomenon under investigation. Alongside this, I consider how pressures op-

erating during learning or comprehension may pull against — or pull with — those arising from production.

1.4.1 Regularity

The claim that languages are "regular" hinges on two key observations: first, that they are governed by systematic rules, and second, that even exceptions to these rules are still systematic in their own way. For example, the default strategy for marking plurality on nouns in English is to add the *-s* suffix (e.g. *dog* → *dogs*). Of course, this is not an exceptionless rule, and so-called irregular nouns are plentiful. Yet irregular nouns are only irregular with respect to the wider system: they are, for the most part, internally consistent (so *mice* are always *mice* and never *mouses* or *meese*). Thus, where variation does exist, it is generally predictable. Random variation — where there are no conditioning factors that govern the choice of one variant over another — is thought to be rare in natural languages (Givón 1985), at least in the output of native speakers (Johnson et al. 1996). But why do languages exhibit this kind of regularity? A vast body of work has highlighted that humans have some kind of expectations about regularity in language, and will tend to introduce it when it doesn't exist: a process known as *regularisation* (e.g. DeGraff 1999; Ferdinand et al. 2019; Hudson Kam & Newport 2005, 2009; Reali & Griffiths 2009; Saldana et al. 2021; Singleton & Newport 2004; K. Smith & Wonnacott 2010).

In Chapter 2, I provide new evidence for a production-based account of regularisation. In an artificial language learning experiment, I show that one version of regularisation — where one variant increases in frequency at the expense of others — arises only when the production task is made harder. Specifically, the way I make the production task harder is by taxing working memory with a concurrent sequence recall task, thus pointing to memory constraints as an underlying mechanism in driving regularisation behaviour. By contrast, making the learning task harder in the same way did not give rise to this kind of regularisation behaviour. However, I also observe another type of regularisation — where variation is maintained, but different variants are specialised for different contexts — and this seems to have less to do with online production effects, and more to do with a learning bias in favour of non-random patterns.

Zooming in on the first of these results, I implement a computational "urn" model — a very simple type of exemplar model (Spike et al. 2017) — which formulates memory constraints during production as a simple self-priming mechanism, whereby variants that are produced more often become increasingly accessible for future productions. I show that this model generates the same pattern of results as the human participants.

1.4.2 Rule learning

There is good evidence that language learning is boosted by engaging in more active production tasks, compared to more passive comprehension tasks. This holds both for infants learning their first language (Bohman et al. 2010; Donnelly & Kidd 2021; Ribot et al. 2018), and for adults learning a second language (Izumi 2002; Keppenne et al. 2021; Swain 2005). For adults, production practice improves both the initial learning of grammatical rules (Hopman & MacDonald 2018) and the ability to generalise these rules to novel items (Hopman 2022). However, it is also known that adult learners do not acquire all kinds of linguistic rules equally well; in particular, morphological rules like case marking pose a significant challenge (e.g. Jordens et al. 1989; Kenanidis et al. 2023; Papadopoulou et al. 2011).

The work presented in Chapter 3 brings these two strands of research together to ask whether production practice can boost learning of morphological rules in particular. I did this project in collaboration with a fellow PhD student, Elizabeth Pankratz, at the intersection of our areas of interest: mine in language production, and hers in rule learning. For my part, I was interested in whether the challenges associated with production could change the kind of structure learners would induce from a new language when more than one analysis was available. We designed an artificial language that used two grammatical rules to cue thematic role: a consistent word order, and case marking morphology. We had participants practise the language with either a more active, production-like task (assembling syllables into sentences) or a more passive, comprehension task (reading sentences and choosing a matching picture). We then tested which rule(s) participants had learned by asking them to judge new sentences. Our idea was that participants who had a chance to actively manipulate the syllables would be more likely to notice a morphological rule that might otherwise have passed

them by. However, although production seemed to improve learning in a very basic sense — we excluded fewer participants from the production group for low accuracy on familiar items — we did not find any evidence that the different tasks affected the kind of rules learners acquired. To preview our discussion, we suspect that this null result has less to do with the utility of production practice and more to do with some shortcomings in our experimental design.

1.4.3 Word similarity

Words of different languages, naturally, sound different from each other. But within a language too, we might intuitively predict that words within a language would be as different from each other as possible to avoid potential confusion. Yet this is not what we see when we look at how sound sequences are distributed within individual languages: some are always vastly more frequent than others. Clearly, some of this skew is due to phonotactic constraints and productive morphology: sound sequences that can occur in more contexts will be more frequent. But even accounting for these factors, cross-linguistic corpus analysis reveals a tendency for lexicons to be more phonetically clustered than would be expected by chance (Dautriche et al. 2017a); that is, words are more similar to each other than they really need to be. However, although this pattern is suggestive about the interplay between different functional pressures that constrain communication, evidence for a causal relationship between lexicon structure and specific communicative mechanisms is lacking.

Chapter 4 aims to fill this gap. I start with a small corpus study which replicates the key finding that lexicons are surprisingly clustered, by benchmarking real words of English against a range of random and phonotactically-controlled baselines. I then develop a cultural evolutionary agent-based model to investigate what mechanisms cause initially-random lexicons to become more clustered through repeated use. I show that natural-language-like levels of clustering emerge from a trade-off between competing communicative pressures: a production-side pressure to re-use more easily articulated sounds, and a comprehension-side pressure for distinctiveness of word-forms. With only one of these pressures at work, the resulting lexicons tend to inhabit an extreme region of the possible design space: production pressures alone give rise to

maximally clustered lexicons, while comprehension pressures alone give rise to maximally disperse lexicons. Finally, I pick up on an observation about how clustering tends to be distributed across real lexicons, whereby more frequent words are more tightly clustered, while lower frequency words are more distinctive (Frauenfelder et al. 1993; King & Wedel 2020; Landauer & Streeter 1973; Mahowald et al. 2018; Meylan & Griffiths 2024). I test whether such frequency effects emerge first in the model, and then in a series of communication experiments with real human participants. Overall, my results lend support only to a weak relationship between frequency and clustering, which depends to some extent on the way I manipulate both frequency itself and the communicative pressures at work. However, in the experiments as in the model, extreme behaviours emerge when only one of these pressures is present, showing again that it is the balancing act between competing pressures that leads to a happy middle ground.

Chapter 2

Working memory and the regularisation of linguistic variation

Tedious, thought it would never end.

Anonymous Prolific participant

Author contributions

The main body of this chapter is an exact reproduction of Keogh et al. (2024), a paper published in *Cognitive Science* in April 2024. The paper was co-authored with my two supervisors, Jennifer Culbertson and Simon Kirby. The hypotheses and experimental design were developed during supervision meetings where all three authors were present and contributing. I created the experiment software, collected the data, conducted the analysis and wrote the first draft of the paper. I also developed the model described in Section 3 independently. Both co-authors provided feedback during the writing and revision of the paper.

Open materials

All materials, code and data used in this chapter are freely available at https://osf.io/9e27b/.

Preamble: An introduction to working memory

In the following paper, I investigate how regularity in language might be shaped by limitations in *working memory*. But before getting into this question, it's worth reviewing some prominent theories and models of working memory, to understand *why* we might expect to see such a relationship.

Working memory refers to the temporary storage and manipulation of very limited amounts of information in active attention — for example, holding a phone number in memory for long enough to write it down, or retaining partial results while solving an equation. The term is often used synonymously with the more general "short-term memory", although there may be some minor distinctions which I will not get into here (e.g. Aben et al. 2012; Cowan 2008). Working memory is a key part of the cognitive architecture supporting human language: it enables us to construct utterance plans ahead of production (MacDonald 2013), and to decode the meaning of incoming linguistic input during comprehension (Lewis et al. 2006).

Models of working memory can be broadly divided into two classes: multi-component (e.g. Baddeley 1992, 2000; Baddeley & Hitch 1974; Jackendoff 2002, 2007; Just & Carpenter 1992; R. C. Martin & Romani 1994; Waters & Caplan 1996) and emergent (e.g. Acheson & MacDonald 2009; Cowan 1993; MacDonald 2016; MacDonald & Christiansen 2002; Majerus 2013; Postle 2006; Schwering & MacDonald 2020). The earliest — and still most dominant — models fall squarely into the multi-component camp. For example, in Baddeley's highly influential model, there is a sharp distinction between passive storage (in "buffers") and active processing (in specialised "slave systems"). Critically, such models also assume that language processing is handled by a working memory system that is functionally separate from the representation of linguistic knowledge in long-term memory (Figure 2.1).

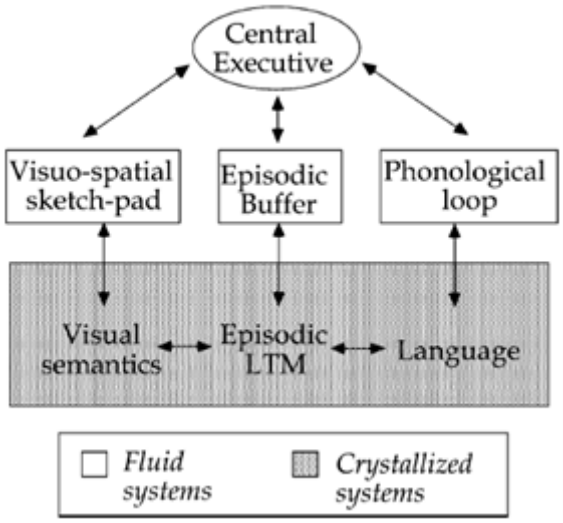

Figure 2.1: The prototypical working memory model (Baddeley 2003). The central executive allocates and retrieves information from other components, while the two slave systems — the phonological loop and the visuospatial sketchpad — are specialised for storing and processing verbally-coded and visually coded information, respectively. The episodic buffer integrates information from the two slave systems and long-term memory and passes it to the central executive.

Emergent accounts, on the other hand, argue that processing capacity emerges from linguistic experience and is not a primitive that can vary independently; on this view, working memory is simply the part of long-term memory that is currently activated. Emergent accounts also do not generally distinguish between storage and processing; indeed, some proponents of this view argue that there is *no* passive storage in working memory, and the *only* way information is maintained is through processing mechanisms like subvocal rehearsal (Buchsbaum & D'Esposito 2019; Postle 2006). More precisely, Acheson and MacDonald (2009) propose that it is the language production architecture which is co-opted for this task; in other words, working memory essentially *is* language production.

Importantly, on both accounts, we should expect the kind of dual-task paradigm I use in the following experiment — where participants are asked to memorise and recall short sequences of digits at the same time as completing a linguistic task — to have a larger effect during production than during learning. In the classic multi-component model (Baddeley & Hitch 1974), the phonological loop itself consists of two sub-components which handle storage and processing respectively: the *phonological store* temporarily holds auditory memory traces, and an *articulatory rehearsal process* refreshes those memory traces as they decay. In emergent models too, rehearsal is the

key mechanism by which verbal information is maintained. The two camps diverge on whether this rehearsal process — sometimes known as "inner voice" — is part of a discrete working memory system, but essentially concur that it is grounded in language production. Therefore, when I ask participants to hold a digit sequence in memory while producing phrases in a target language, the cognitive resources they have available for language production are split between the two tasks. Conversely, when I ask them to hold a digit sequence in memory while passively observing phrases to be learned in the target language, it is not clear that the same cognitive resources are allocated to the two tasks. Indeed, this is fundamentally why my hypotheses in the experiment are about production-side explanations for regularisation. I include the LEARNING LOAD conditions primarily as a control, and by way of acknowledging that some previous work (e.g. Elman 1993; Goldowsky & Newport 1993; Pitts Cochran et al. 1999) has suggested a role for working memory limitations in language acquisition (albeit controversially: see Rohde and Plaut 2003).

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13435

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13435

Predictability and Variation in Language Are Differentially Affected by Learning and Production

Aislinn Keogh, Simon Kirby, Jennifer Culbertson

Centre for Language Evolution, University of Edinburgh

Received 11 August 2023; received in revised form 1 March 2024; accepted 6 March 2024

Abstract

General principles of human cognition can help to explain why languages are more likely to have certain characteristics than others: structures that are difficult to process or produce will tend to be lost over time. One aspect of cognition that is implicated in language use is working memory—the component of short-term memory used for temporary storage and manipulation of information. In this study, we consider the relationship between working memory and regularization of linguistic variation. Regularization is a well-documented process whereby languages become less variable (on some dimension) over time. This process has been argued to be driven by the behavior of individual language users, but the specific mechanism is not agreed upon. Here, we use an artificial language learning experiment to investigate whether limitations in working memory during either language learning or language production drive regularization behavior. We find that taxing working memory during production results in the loss of all types of variation, but the process by which random variation becomes more predictable is better explained by learning biases. A computational model offers a potential explanation for the production effect using a simple self-priming mechanism.

Keywords: Working memory; Language evolution; Artificial language learning; Regularization; Language production; Urn model

Correspondence should be sent to Aislinn Keogh, Centre for Language Evolution, University of Edinburgh, Edinburgh, EH8 9AD, UK. E-mail: aislinn.keogh@ed.ac.uk

1. Introduction

Language is created in real time: successful processing requires us to rapidly turn complex input into the correct mental representations, while successful production requires us to rapidly turn our mental representations into meaningful output. However, the finite nature of human memory imposes a bottleneck on these processes, shaping the kinds of structures that can persist as languages evolve (Christiansen & Chater, 2016, 2008; Futrell, Mahowald, & Gibson, 2015; Kirby, 1999; MacDonald, 2013). It has long been acknowledged that working memory—the component of short-term memory used for temporary storage and manipulation of information (including linguistic information)—is severely limited in its capacity (Baddeley & Hitch, 1974; Baddeley, 2000; Cowan, 2001; Gobet & Clarkson, 2004; Miller, 1956). Cognitive constraints such as these can help to explain why languages look the way they do: as languages are passed from person to person, properties that make them easier to process or produce are likely to edge out those that place a more significant burden on working memory. Thus, some processes of language change might arise as a result of an interaction between linguistic representations and constraints on memory and other general principles of human cognition (Culbertson & Kirby, 2016).

In this study, we consider the role of working memory limitations in the regularization of linguistic variation. Regularization is a well-documented process of language change whereby a language becomes less variable (on some dimension) over generations. This process has been argued to be driven by individual language learners and users, who produce output that is less variable than their input (Hudson Kam & Newport, 2009). Repeated across many individuals and generations, this behavior is one way in which emerging languages may acquire systematic rules and regularities (Smith & Wonnacott, 2010). For example, nouns in English generally mark plurality with the regular *-(e)s* suffix (e.g., *dog* → *dogs*), but even among irregular nouns there are identifiable, semi-productive patterns (e.g., the vowel change in *mouse* → *mice* and *louse* → *lice*, or null marking in *fish* and *sheep*). Furthermore, while there is considerable variation in the English plural system overall, the choice of form for any given word is generally phonologically or lexically conditioned. By contrast, random variation—where there are no conditioning factors—is rare in natural languages (Givón, 1985), at least in the output of native speakers (Johnson, Shenkman, Newport, & Medin, 1996). Thus, while variation is ubiquitous, it tends to be predictable in some way.

1.1. Regularization of unpredictable variation

There is a wealth of evidence that language users reduce unpredictable variation, both in the lab and in natural language. Children exposed to unpredictable variation in artificial language learning studies tend to regularize at a system-wide level, increasing their use of one variant (usually the form they encountered most frequently in the input) to the exclusion of others (Hudson Kam & Newport, 2005, 2009; Schwab, Lew-Williams, & Goldberg, 2018). This behavior persists even when the most frequent form in the input is not actually very frequent at all (Austin, Schuler, Furlong, & Newport, 2022). Regularization behavior can also be observed in adults, although potentially to a lesser degree or in a narrower range

of circumstances than in children (Culbertson & Newport, 2015; Hudson Kam & Newport, 2009). For example, adults regularize more when the number of alternating variants increases (Ferdinand, Kirby, & Smith, 2019; Hudson Kam & Newport, 2009; Saldana, Smith, Kirby, & Culbertson, 2021), when generalizing to novel contexts (Wonnacott & Newport, 2005), and when attempting to coordinate with other individuals in communicative tasks (Fehér, Ritt, & Smith, 2019; Fehér, Wonnacott, & Smith, 2016; Kamps, Ferdinand, & Kirby, 2014; Perfors, 2016). Furthermore, even when adults maintain variation, they often still regularize at a lower level, making variation more predictable by conditioning it on some aspect of the context like lexical item or grammatical category (Samara, Smith, Brown, & Wonnacott, 2017; Smith & Wonnacott, 2010). And although individual adults may show weaker evidence of regularization than children, this effect may nevertheless be amplified through cultural transmission as small increases in regularity accumulate over generations (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Smith et al., 2017).

In natural language, regularization of unpredictable variation has been observed in deaf children exposed to inconsistent linguistic input, both in the acquisition of existing signed languages from non-native users (Singleton & Newport, 2004) and in the formation of new signed languages (Senghas, Coppola, Newport, & Supalla, 1997; Senghas & Coppola, 2001). Regularization has also been argued to be at play in the emergence of stable creole languages from highly variable pidgin languages (Aitchison, 1996; Bickerton, 1981; DeGraff, 1999; Siegel, 2007).

1.2. Regularization of predictable variation?

It is less clear whether the cognitive mechanisms driving regularization act as strongly on predictable patterns of variation. In natural language, while there are certainly cases of irregular forms (e.g., *cow* → *kine* in Middle English) shifting to the regular pattern, there is some evidence that *irregularization* is roughly as prevalent a process as regularization, and that the main driver of increased regularity is the introduction of new lexical items (which tend to be regular) rather than the regularization of existing items (Cuskley et al., 2014). Furthermore, regularization is highly frequency-dependent: high-frequency forms tend to exhibit stable irregularity, while lower frequency forms are more likely to regularize (Carroll, Svare, & Salmons, 2013; Cuskley et al., 2017; Lieberman, Michel, Jackson, Tang, & Nowak, 2007; Smith, Ashton, & Sims-Williams, 2023).

Artificial language learning experiments testing the acquisition of conditioned variation also provide somewhat mixed evidence. Although this kind of variation is clearly far more typical of natural language than the unpredictable variation usually targeted by regularization experiments, it is not always learned or reproduced more accurately. When these patterns of variation are only probabilistic, children can struggle, whether conditioning is by linguistic features like syntactic role (Hudson Kam, 2015) or by salient semantic features like natural gender (Schwab et al., 2018). However, children *are* sensitive to certain conditioning cues (especially phonological: Culbertson, Jarvinen, Haggarty, & Smith, 2019; Karmiloff-Smith, 1981; Pérez-Pereira, 1991; Gagliardi & Lidz, 2014) and seem to regularize less (or not at all) when conditioning is deterministic (Austin et al., 2022; Brown, Smith, Samara, &

Wonnacott, 2022; Samara et al., 2017; Wonnacott, 2011). Adults generally have less difficulty acquiring conditioned variation—either probabilistic (Schwab et al., 2018) or deterministic (Austin et al., 2022; Hudson Kam & Newport, 2009)—and often maintain this kind of variation across multiple simulated generations in iterated learning experiments (Smith et al., 2017, Smith et al., 2023; Smith & Wonnacott, 2010). However, as with children, adults' performance varies according to the presence or salience of conditioning cues: neither age group appears to readily acquire arbitrary subclass distinctions (Braine et al., 1990; Culbertson & Wilson, 2013; Frigo & McDonald, 1998; Smith, 1969).

Overall then, there seems to be good reason to suspect that at least certain kinds of conditioned variation will also be regularized—although seemingly to a lesser extent than unpredictable variation.

1.3. What causes regularization?

Whether regularization should target all kinds of variation—or only unpredictable variation—might depend on the underlying cause of the behavior. However, the specific mechanism driving regularization is not agreed upon.

One possibility is that regularization arises from a failure to encode variation during learning (Culbertson, Smolensky, & Wilson, 2013; Hudson Kam & Newport, 2009). In other words, when individuals produce a more regular language than the one they were exposed to, they may be faithfully producing what they remember of their input. On this account, age differences in regularization behavior might be explained by developmental changes in general learning mechanisms; perhaps, by not acquiring the full complexity of their input, children are better able than adults to extract regularities from noise (Hudson Kam & Newport, 2009; Rische & Komarova, 2016). However, tasks that provide a more direct window on individuals' internal representations (e.g., grammaticality judgments or frequency reports) provide evidence that even those who exhibit the most extreme regularization behavior still show awareness of the inconsistencies in their input, including for very complex patterns (Austin et al., 2022; Ferdinand et al., 2019; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2009; Schwab et al., 2018; Saldana et al., 2021). Furthermore, Perfors (2012) found that requiring participants to attend to a secondary task while they learn an artificial language impaired vocabulary acquisition, but had no effect on the strength of regularization behavior, suggesting that regularization is not an inevitable consequence of imperfect learning.

This suggests that regularization may be primarily a production-side process. However, this still leaves open several possible mechanisms. For example, regularization in production may be driven by specific pragmatic contexts. In line with this, adults seem to regularize more when they understand that the variation in their input is genuinely random (Perfors, 2016), suggesting that when they maintain variation, it is because they think it is meaningful (Clark, 1988). Regularization behavior is also stronger during communicative tasks, either due to accommodation between interlocutors or because individuals strategically remove aspects of the linguistic signal that do not correlate with differences in meaning to maximize communicative success. (Fehér et al., 2016; Fehér et al., 2019). The pragmatic account straightforwardly predicts that unpredictable variation will be regularized, but it is not clear that these

mechanisms would also target predictable variation. Conditioned variation already satisfies language users' expectation that variation in language should be rule-governed (Wonnacott & Newport, 2005), so getting rid of it would not obviously increase communicative success; in fact, failing to observe the rules of the language in this way might even *hinder* communication. Accommodation between interlocutors too would presumably favor the lexically specific rules that both had acquired.

Alternatively, there may be purely cognitive factors that drive regularization in production, such as working memory limitations. One hypothesis which has received some experimental support is that regularization arises from limitations on memory retrieval during language production (Hudson Kam, 2019; Hudson Kam and Chang, 2009). The exact mechanism is unclear, but one possibility is that, when retrieval is difficult, variants that have been produced recently become increasingly accessible for retrieval on subsequent productions through repetition priming (Hudson Kam, 2019; Schwab et al., 2018). These ideas are consistent with models in which language production is not simply a perfect reflection of what has been learned but is also constrained by online demands like ease of retrieval (Goldberg & Ferreira, 2022; MacDonald, 2013). On such an account, we might expect that regularization would target both predictable *and* unpredictable variation since an overall higher frequency form might be more easily retrieved in general, even if specific lexical items had been encountered in different constructions.

Several previous studies suggest that memory retrieval is a factor in driving the regularization of *unpredictable* variation. On the one hand, this hypothesis predicts less regularization when retrieval is less taxing. Indeed, Hudson Kam and Chang (2009) found that adults more closely matched the statistics of their input when the production task was made easier. Similar results have been found with children, who seem to regularize less when the burden of lexical access is eased through the use of English nouns in semi-artificial languages (Samara et al., 2017; Wonnacott, 2011). Another way of getting at the question is to directly interfere with working memory by asking participants to attend to multiple tasks simultaneously. This method aims to disrupt a specific aspect of linguistic working memory—either encoding or retrieval, depending on when it is administered—in order to provide evidence for its involvement. Perfors (2012) performed such a manipulation during learning which, in line with the production-side account, did not result in increased regularization. Hudson Kam (2019) replicated this result with a much more complex language and offered some preliminary evidence that a comparable manipulation during production *may* contribute to increased regularization. Specifically, participants subject to interference during production seemed more likely to regularize on an item-by-item basis (i.e., condition their use of different variants on lexical items).

1.4. The present study

In this paper, we further explore the role of working memory (and memory retrieval) in driving regularization of both predictable and unpredictable variation. In line with Perfors (2012), our goal is to look for evidence of regularization in a simple language which isolates the phenomenon of interest and removes superfluous elements like word order, transitivity,

and negation that are present in the language of Hudson Kam (2019). However, in common with Hudson Kam (2019), we ask whether interfering with working memory during language *production* (rather than learning) leads to regularization. Additionally, we ask whether this production-side mechanism targets predictable variation to the same extent as unpredictable.

To preview, we provide experimental evidence that regularization of both predictable and unpredictable variation does indeed arise under memory load during production. Interestingly, we also find that working memory limitations have some effect on regularization during learning, contrary to previous studies. Finally, we implement a computational model of regularization in production via a simple self-priming mechanism by which a high-frequency variant becomes increasingly accessible for retrieval through repeated production.

2. Experiment

We use a 2×3 between-subjects design to investigate the effect of memory limitations on the regularization of linguistic variation in six experimental conditions. We trained participants on an artificial language exhibiting variation in nominal marking that was either probabilistically lexically conditioned (PREDICTABLE conditions) or random (UNPREDICTABLE conditions). We then tested participants' ability to produce *noun* + *marker* combinations in the language, and their ability to estimate the frequency with which particular *noun* + *marker* combinations had appeared in the input (a measure of learning, following previous work, e.g., Ferdinand et al., 2019). We used an interference task, modeled after the concurrent load tasks used by Perfors (2012) and Hudson Kam (2019), to tax working memory during either learning (LEARNING LOAD conditions) or production (PRODUCTION LOAD conditions); in a third, baseline condition, there was no such task (NO LOAD conditions).

In line with the production-side account of regularization, we predicted that participants would *produce* a more regular language than the one they learned, regardless of the type of variation (predictable or unpredictable). By contrast, we predicted no regularization in participants' frequency estimates. In line with the memory retrieval hypothesis, we predicted that we would see the clearest evidence for reduction of variation when taxing working memory during production. Finally, to test our hypothesis about the relationship between predictable and unpredictable languages, we predicted that the effect of memory limitations during production would be modulated by variation type, with greater regularization of unpredictable languages.

2.1. Methods

The study was approved by the PPLS Ethics Committee at the University of Edinburgh and was pre-registered with the Open Science Foundation (https://osf.io/vqyej).

2.1.1. Participants

We recruited 220 participants via Prolific. Participants were adult, self-reported native English speakers with no known language disorders. They were provided with a downloadable information sheet and gave informed consent to participate. The experiment took around

Table 1
Number of participants per condition submitted to analysis

	Predictable	Unpredictable
No load	29	28
Learning load	30	28
Production load	29	29

Table 2
Distribution of plural markers (P_i) across nouns (N_j) in the two variation conditions

(a) Predictable Input Languages

	N1	N2	N3	N4	N5	N6	Total
P1	7	7	7	7	1	1	30
P2	1	1	1	1	7	7	18
Total	8	8	8	8	8	8	48

(b) Unpredictable Input Languages

	N1	N2	N3	N4	N5	N6	Total
P1	5	5	5	5	5	5	30
P2	3	3	3	3	3	3	18
Total	8	8	8	8	8	8	48

20 minutes to complete ($M = 18.01$, $SD = 8.48$), for which participants were paid £3 (above the UK national minimum wage). Forty-seven participants were excluded for the following pre-registered reasons: self-reporting the use of written notes in an exit questionnaire contrary to instructions (three), data saving errors (one), failing to provide usable data on more than two critical trials (38),[1] and button mashing (five).[2] This left us with data from 173 participants (Table 1).

2.1.2. Materials

The artificial language consisted of orthographically presented labels paired with six images. Each image depicted a pair of animals and was described by a two-word label: one word for the noun and one word indicating plurality (presented in the English frame "Here are two..."). Noun labels were designed to be similar to English onomatopoeia (e.g., "buzzo" for a bee) to ensure that learning of this part of the label would be trivially easy for all participants, regardless of memory load. Nouns were paired with one of two plural markers, both non-English CVC monosyllables ("mej" and "huv"). The mapping of nouns to plural markers varied according to condition (Table 2). In PREDICTABLE conditions, the choice of one plural or the other was probabilistically conditioned on the noun. Four nouns were randomly assigned to one plural marker (the "regulars") and two to the other marker (the "irregulars"). A small amount of noise was then added to this mapping, such that, for n repetitions of a given noun in the training set, that noun appeared with its assigned plural marker $n - 1$ times (87.5%) and once with the other marker (12.5%). This noisy conditioning meant that

participants could regularize without having to produce a description they had never observed. In UNPREDICTABLE conditions, plural markers varied randomly across nouns with no conditioning: all nouns appeared with one marker 62.5% of the time and with the other 37.5% of the time. Both markers appeared with the same overall frequency in the two variation conditions, allowing us to assess the extent to which item-specific patterns affect the tendency to regularize, even when the global language statistics are identical.

2.1.3. Procedure

The experiment was written in JavaScript using the JsPsych library (de Leeuw, 2015) and ran in participants' web browser. Participants were randomly assigned to one of the six conditions at the start of the experiment. The experiment consisted of three phases: training, production, and estimation.

In the training phase, participants were asked to learn the words used to describe the animals. Each of the six images was shown eight times for a total of 48 trials. The order of presentation was randomized. On each training trial, an image was presented for 1000 ms and then a description of the form "Here are two *noun* + *plural*" appeared below the image. The image and description disappeared after 3000 ms and participants clicked a "continue" button to advance to the next trial.

In the production phase, participants were asked to produce descriptions for the same set of stimuli. Again, each of the six images was shown eight times for a total of 48 trials.[3] On each production trial, participants saw an image and a partial description, consisting of an English frame and two gaps for the artificial words: "Here are two _____ _____". They were asked to fill in the gaps by clicking two buttons from an array consisting of all nouns and plural markers in the language. This multiple-choice production task is intended to simulate the process of a fluent speaker selecting words from a stably represented mental lexicon. It allows us to observe the effects of online demands in production while minimizing the possibility that participants' choice of words is driven by incomplete learning.[4] Buttons were blocked into nouns (on the left) and plural markers (on the right), with the order of buttons randomized within each block and a clear gap between blocks. However, participants were not forced to click one button from the first block and one from the second. No feedback was provided; participants simply saw the gaps filled with whichever words they had selected. The full label they had assembled was displayed for 1000 ms before they advanced to the next trial.

Finally, in the estimation phase, participants were asked to estimate how often they had seen each noun with each plural marker in training. All six images appeared in a random order on one page, each accompanied by a continuous slider over percentages. All sliders started in the middle, and participants were required to move every slider before they could advance. Each slider had three labels: "always *P1*" at 0%, "equal *P1/P2*" at 50%, and "always *P2*" at 100%. The assignment of plural markers to the two ends of the slider was randomized for each participant, but identical for all sliders.

In LEARNING LOAD and PRODUCTION LOAD conditions, participants were told that we were interested in how well people can learn or produce (respectively) a new language when the task is difficult, so they would also be asked to memorize and recall short sequences of numbers alongside the main task. They were told that they would be given feedback throughout

on their performance on this task. The aim was to occupy participants' conscious attention with the secondary task to disrupt the part of working memory they would otherwise have devoted to the linguistic task. The task was sandwiched around (i.e., concurrent with) each trial in either the training phase (LEARNING LOAD) or production phase (PRODUCTION LOAD). First, a pseudorandom sequence of three digits was displayed for 2500 ms and participants were asked to memorize the numbers in order. A new sequence was generated on each trial by sampling the set of digits 0–9 without replacement, with the constraint that each digit n was never neighbored on either side by $n + 1$ or $n - 1$, preventing any obvious patterns appearing in the sequences that might have made them easier to remember. Participants then completed the main training or production trial. Immediately following this, participants were asked to retype the numbers they had just memorized, in order. They were given feedback on the number of digits they had recalled in the correct position and how long they had taken to respond, to encourage both speed and accuracy.

A schematic of the experimental procedure for the PRODUCTION LOAD conditions is given in Fig. 1.

2.1.4. Analysis

We take an information theoretic approach (Shannon, 1948) to quantifying variation and regularization (following, e.g., Ferdinand et al., 2019; Perfors, 2016; Samara et al., 2017; Smith & Wonnacott, 2010). This analytic approach is sensitive even to small changes in frequency distributions, regardless of whether those changes are in the direction predicted by the input (i.e., even if participants regularize with the minority variant[5]). We report three specific measures below: entropy, conditional entropy, and mutual information (MI). The first two measures were pre-registered, the third is an addition which we explain below.

Entropy: The total amount of variability in a plural marking system is captured by the entropy of the frequency distribution of plural markers across the language. Taking plural marking as a discrete random variable V with possible variants $v_1 \ldots v_n$ which occur with probability $p(v_1) \ldots p(v_n)$, the entropy of a language is given as

$$H(V) = -\sum_{v_i \in V} p(v_i) \log_2 p(v_i).$$

More skewed distributions (i.e., languages in which one plural marker is used more frequently) exhibit lower entropy. A maximally regular language (with only one plural marker) would score 0, while a maximally irregular language (where both markers appear 50% of the time) would score 1. Since the frequency distribution of plural markers across the input languages in both PREDICTABLE and UNPREDICTABLE conditions is identical, the languages are matched for entropy (0.95 bits).

Conditional entropy: The predictability of a plural marking system can be measured by considering how variable individual nouns are: a language where each noun only uses one plural marker is more predictable than one where nouns can take any marker. The average variability of individual nouns in a language is captured by the conditional entropy of the

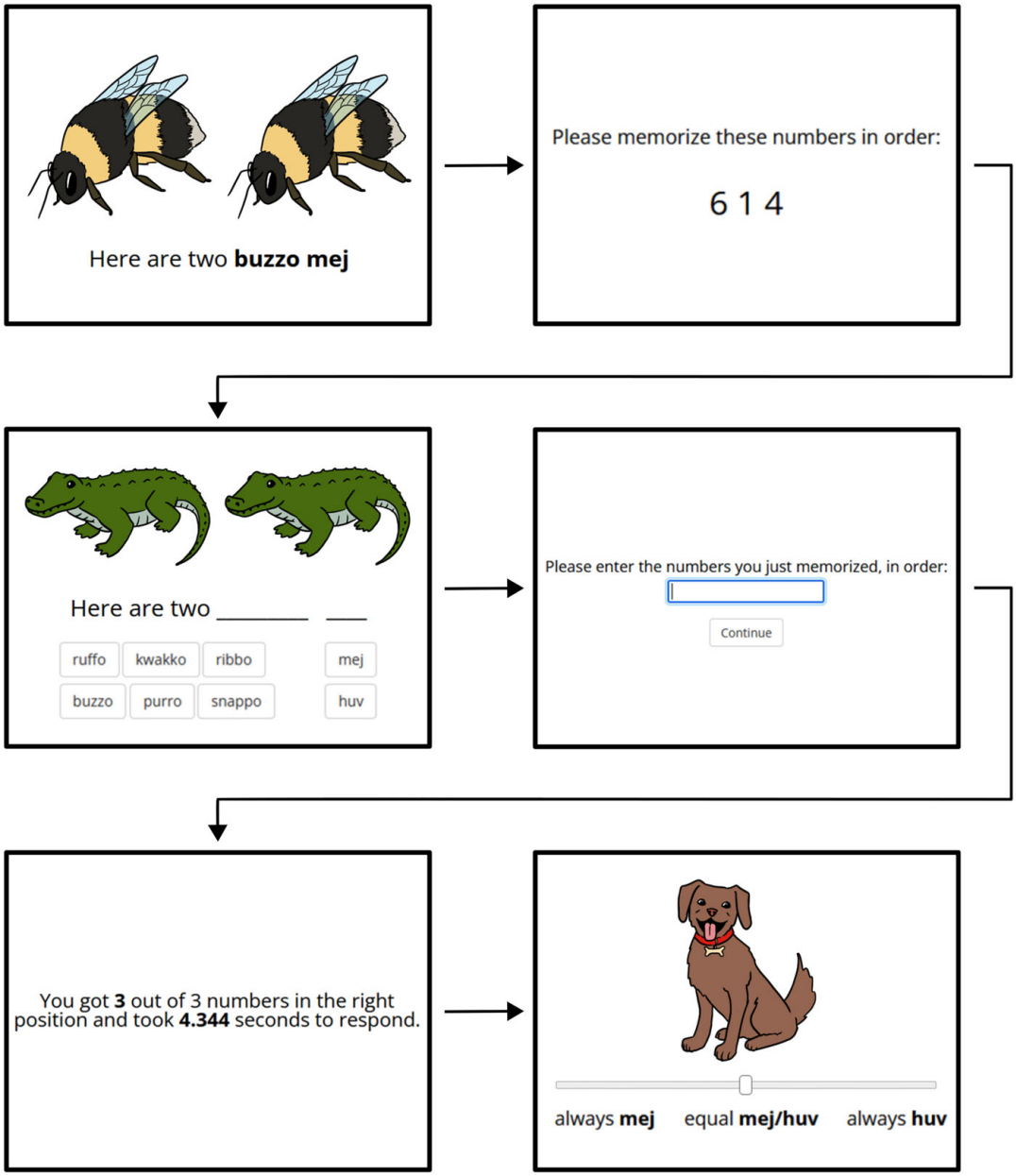

Fig. 1. Schematic of the experiment: PRODUCTION LOAD condition. Top to bottom, following arrows: training trial, digit sequence presentation, production trial, digit sequence recall, feedback, and estimation trial. Participants in LEARNING LOAD conditions would instead have seen the digit sequence presentation and recall trials sandwiched around each training trial. Participants in NO LOAD conditions would not have seen these digit sequence trials.

frequency distribution of plural markers, given the noun being marked. Given a set of variants V (plural markers) and a set of contexts in which these variants appear C (nouns), the conditional entropy of a language is given as

$$H(V|C) = -\sum_{c_j \in C} p(c_j) \sum_{v_i \in V} p(v_i|c_j) \log_2 p(v_i|cj).$$

The variability of individual nouns in UNPREDICTABLE input languages mirrors that of the language as a whole, so entropy and conditional entropy are matched for these languages (0.95 bits on both measures). On the other hand, PREDICTABLE input languages have lower conditional entropy since individual nouns in these languages are less variable than the language as a whole (0.54 bits).

Mutual information: When either entropy or conditional entropy decreases, we can infer that the language has become more regular in some sense. However, here we would like to distinguish between regularization at the lexical level (i.e., a given plural marker used more with a particular noun) and regularization across the language as a whole (i.e., a given plural marker used more often overall). Conditional entropy does not allow us to do this since it is affected by overall entropy: when a language becomes less variable overall, the choice of plural marker necessarily becomes more predictable. We, therefore, added a third measure to our set of pre-registered variables: mutual information (MI). MI is the difference between the two entropy measures[6] and allows us to isolate the amount of predictability that is specifically explained by lexical conditioning. MI of 0 indicates a complete absence of lexical conditioning; this is the case both when there is no variability (since there is nothing to condition here), and when the variability of individual nouns mirrors that of the language overall (as in the UNPREDICTABLE input languages). MI of 1 would indicate that the language as a whole is maximally variable (i.e., the two plural markers are equally frequent overall), but each noun is perfectly non-variable. PREDICTABLE input languages here score 0.41, reflecting the presence of imperfect conditioning in a skewed overall frequency distribution.[7]

2.2. Experiment results

We analyzed the data in R (R Core Team, 2022). Each of the measures described in Section 2.1.4 was calculated for the languages participants were trained on, the languages they produced, and the languages described by their estimates. We investigate regularization as a function of learning by comparing participants' estimates to their training data. We investigate regularization as a function of production by comparing participants' productions to their training data and to their estimates. The dependent variable in all analyses is, therefore, the *change* in the given measure. We define regularization as a reliable *decrease* in entropy or a reliable *increase* in MI. Plots in this section show population-level data; individual-level data are available in Appendix A.

2.2.1. Pre-requisites

In order to test the hypotheses of interest, it is crucial that we first rule out the possibility that any differences between conditions are driven by differences in vocabulary learning or in performance on the interference task. The following mixed effects models were generated using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) and include fixed effects of variation type and memory load, and their interaction, as well as by-participant and by-item random intercepts.

Performance on the interference task was close to ceiling across conditions (overall, $M = 2.84$, $SD = 0.57$). Since the distribution of scores is very left-skewed, we take as our dependent variable the *error rate* (calculated as 3—the number of correct digits), which approximates a Poisson distribution. We performed a mixed effects Poisson regression predicting the error rate by condition. Model comparison revealed that neither variation type ($\chi^2(2, 6) = 0.258$, $p = .879$) nor memory load ($\chi^2(2, 6) = 2.462$, $p = .292$) were significant predictors of performance. These results indicate that participants in all conditions were attending equally well to this task. Furthermore, the level of performance indicates that participants took the task seriously; we can, therefore, be confident that participants did not focus on the main task to the exclusion of the interference task, which would obscure any possible effects in the load conditions.

Noun learning was also close to ceiling across conditions (overall, $M = 0.97$, $SD = 0.17$). We performed a mixed effects logistic regression predicting the log-likelihood of a correct response by condition. Model comparison revealed that neither variation type ($\chi^2(3, 8) = 1.278$, $p = .734$) nor memory load ($\chi^2(4, 8) = 2.719$, $p = .606$) were significant predictors of noun learning. These results indicate that participants in all conditions learned the lexicon equally well.

In summary, any differences in regularization we see across conditions are not due to accidental differences in performance on the memory load task or noun learning.

2.2.2. Main analysis

Inspection of the models specified in our pre-registration revealed that residuals were significantly non-normally distributed (confirmed by Shapiro–Wilk tests) and had non-constant variance over groups (confirmed by Breusch–Pagan tests for heteroscedasticity). Since our data did not meet the assumptions for a linear modeling analysis, the analyses we present here instead evaluate our pre-registered predictions using a simulation-based approach.[8]

Our null hypothesis is that participants' responses reflect a probability-matching strategy (e.g., Estes, 1976; Gardner, 1957; Hudson Kam & Newport, 2005). To determine how much we can expect entropy and MI to change under this strategy, we simulate participants who produce the majority marker for any given noun on any given trial with a probability equal to its frequency in the input. We generate 10,000 runs of 30 such participants and calculate the mean of each run. This gives us a distribution of expected means under the null hypothesis against which we can z-score our real by-condition means. A z-score of < -1.96 indicates a reliable decrease in entropy, while a z-score of > 1.96 indicates a reliable increase in MI.[9]

To identify main effects of our predictors, we take a permutation-based approach. The null hypothesis is that different conditions do not give rise to substantially different behavior. We can generate data that meets this assumption by randomly shuffling the labels for one predictor in our real data. For example, to test for a main effect of variation type, we shuffle the column containing the PREDICTABLE/UNPREDICTABLE labels, thus breaking the association between each data point and its condition label. We carry out this shuffling 10,000 times, calculating the difference between condition means (in the example case, between the mean of all PREDICTABLE and all UNPREDICTABLE conditions) for each run, to give us a distribution of expected differences between conditions under the null hypothesis, against which we can

z-score our real difference.[10] A z-score of > 1.96 indicates that the observed difference between conditions is reliably greater than would be expected by chance.

This permutation analysis also allows us to identify interactions between predictors. The null hypothesis here is that the difference between the levels of one predictor is the same across the levels of the other predictor, that is, the effect of memory load does not depend on variation type or vice versa. Again, we can generate data that meets this assumption by randomly shuffling the labels for one predictor in our real data. For example, to test whether the effect of the PRODUCTION LOAD manipulation differs between variation types, we first shuffle the column containing the PREDICTABLE/UNPREDICTABLE labels then calculate the difference between the PRODUCTION LOAD condition and other memory load conditions (collapsed) separately for the PREDICTABLE and UNPREDICTABLE conditions, and finally calculate the difference between these differences. We carry out this shuffling 10,000 times to generate a distribution of expected differences in differences under the null hypothesis, against which we can z-score our real difference in differences. A z-score of > 1.96 indicates that the observed difference in differences is reliably greater than would be expected by chance.

We can also calculate p-values for all reported statistics by counting the number of values in the relevant null distribution that are as or more extreme than our observed value and dividing this by the number of runs (10,000). Due to the finite nature of the sample, this sometimes gives a value of exactly 0 or 1; in this case, we report $p < .001$ or $p > .999$.

Regularization during learning: We predicted that participants across conditions would show no evidence of having learned a more regular language than the one they were trained on. The estimation task results allow us to assess this prediction. The comparison of interest is thus between the languages participants were trained on and the ones described by their estimates.

Fig. 2a shows the change in entropy. In line with our prediction, we found no reliable decrease in entropy: no condition mean falls below the lower tail of the corresponding null distribution. However, as Fig. 2a shows, there was an increase in MI between the languages participants in UNPREDICTABLE conditions were trained on and the ones described by their estimates: the mean of each of these conditions is well above the null distribution. Permutation analysis confirms a main effect of variation type ($Z = 5.498$, $p < .001$).

To summarize, these results show, in line with our prediction, that the learning process does not drive regularization at a system-wide level: participants are able to encode the overall frequency of different variants in their input. However, we do see evidence of a learning bias for regularization at the lexical level, with learners in the UNPREDICTABLE conditions inferring a pattern of conditioning when no such pattern exists in their input.

Regularization during production: Before analyzing participants' production data, we got rid of trials where the label produced was of an invalid form (i.e., anything other than *noun* + *plural*) or where the noun was incorrect.[11]

Recall that we predicted that taxing working memory during production would lead to greater regularization behavior. We also predicted that we would see greater regularization of unpredictable languages and that this factor would modulate the size of the effect of memory

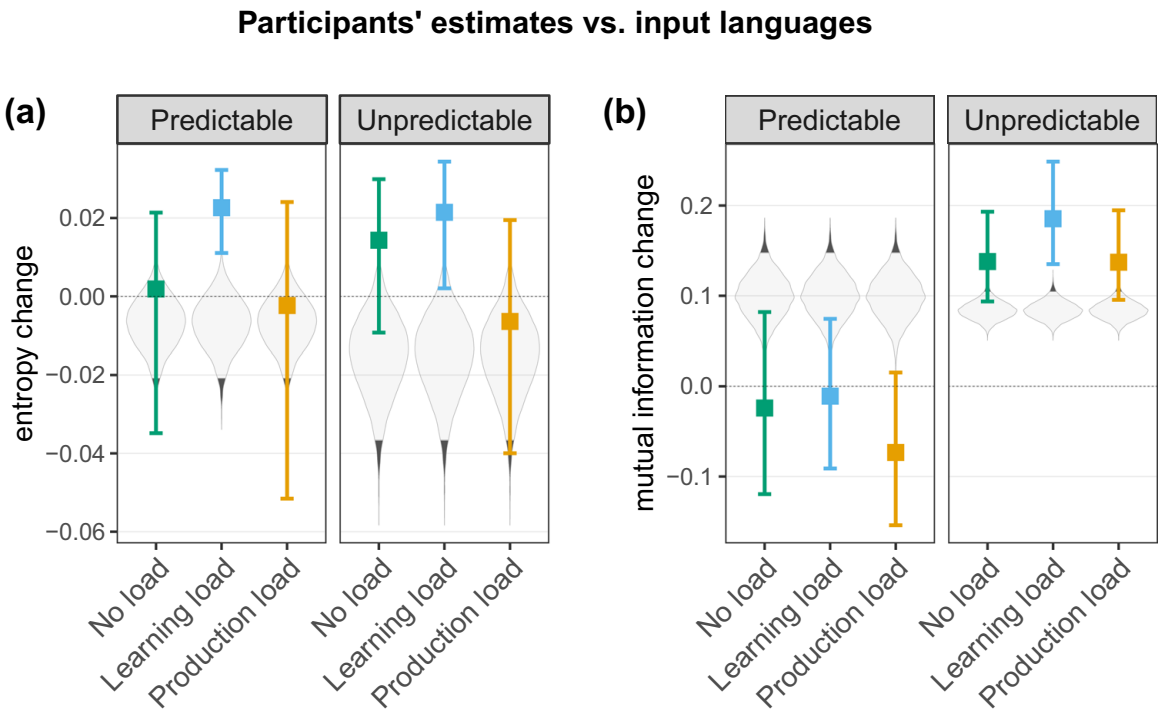

Fig. 2. Change in entropy (left) and mutual information (right) between the languages participants were trained on and the ones described by their estimates, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability-matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. There is no reliable decrease in entropy in any condition, indicating that participants did not underestimate the total amount of variation in their input. However, there is a reliable increase in MI between the languages participants in UNPREDICTABLE conditions were trained on and the ones described by their estimates, indicating that participants in these conditions overestimated the degree of lexical conditioning present in their input.

limitations. To assess these predictions, the comparison of interest is between the languages participants were trained on and the ones they produced.

Fig. 3a shows the change in entropy. In line with the first part of our prediction, the only place we see a reliable drop-in entropy is the PRODUCTION LOAD conditions: the means of these conditions (and no others) are both below the lower tail of the null distributions. Permutation analysis confirms a main effect of memory load, with greater entropy drop in PRODUCTION LOAD conditions than other memory load conditions ($Z = -3.034$, $p = .001$). Contrary to our prediction, permutation analysis reveals no main effect of variation type ($Z = 0.620$, $p = .733$). Although, descriptively, entropy does drop more in the UNPREDICTABLE/PRODUCTION LOAD condition ($M = -0.118$) than in the PREDICTABLE/PRODUCTION LOAD condition ($M = -0.060$), we find no statistical evidence that the effect of the production load manipulation is stronger in the UNPREDICTABLE condition ($Z = 1.092$, $p = .856$). In other words, there is no reliable interaction between variation type and memory load.

As shown in Fig. 3b, we observed an increase in MI across all conditions apart from PREDICTABLE/NO LOAD ($Z = 1.506$, $p = .065$) and PREDICTABLE/PRODUCTION LOAD ($Z = -2.650$, $p = .996$). On this measure, our data, therefore, suggest that there is a general

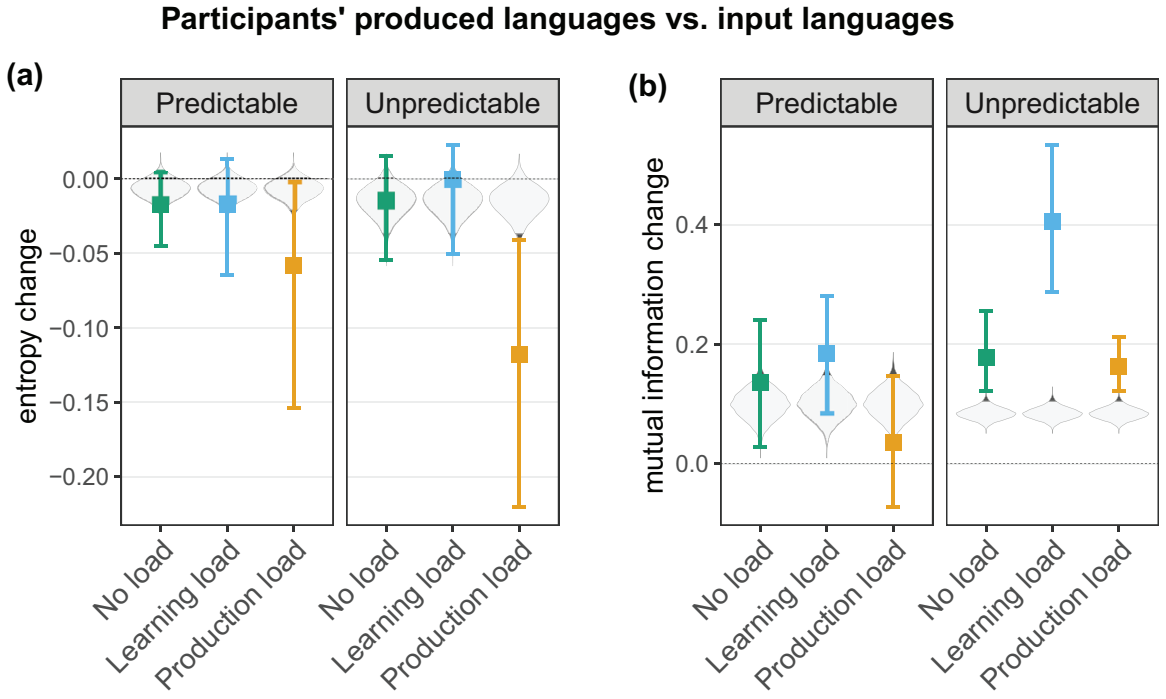

Fig. 3. Change in entropy (left) and mutual information (right) between the languages participants were trained on and the ones they produced, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. Entropy decreases only in PRODUCTION LOAD conditions, indicating that taxing working memory during production increases participants' tendency to over-produce one variant relative to its frequency in the input. MI, on the other hand, increases in all but the PREDICTABLE/NO LOAD and PREDICTABLE/PRODUCTION conditions, and especially so in the UNPREDICTABLE/LEARNING LOAD condition. This seems to reflect a general preference to produce lexically conditioned variation, amplified by memory limitations during learning.

tendency to introduce or boost lexical conditioning, not arising from the same memory mechanism that leads to entropy drop. In line with our prediction, permutation analysis confirms a main effect of variation type, with a greater increase in MI in UNPREDICTABLE conditions ($Z = 2.999$, $p = .002$).[12] Permutation analysis also reveals a main effect of memory load. However, as suggested by Fig. 3b, this is in the opposite direction than predicted: MI increases *less* in PRODUCTION LOAD conditions than other memory load conditions ($Z = -2.745$, $p = .002$). Since inspection of the means suggests that MI actually increased more in LEARNING LOAD conditions, we carried out an exploratory analysis by collapsing NO LOAD and PRODUCTION LOAD conditions together. Permutation analysis on this coding scheme supports the notion that MI increases significantly more in LEARNING LOAD conditions than others ($Z = 3.596$, $p < .001$), suggesting that the preference for lexical conditioning is amplified by memory limitations during *learning*. The interaction analysis we ran for the entropy data is clearly not warranted by the MI data since the main effect does not go in the predicted direction. We carried out a further exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions, but this analysis revealed no reliable interaction between variation type and memory load ($Z = -1.407$, $p = .081$).

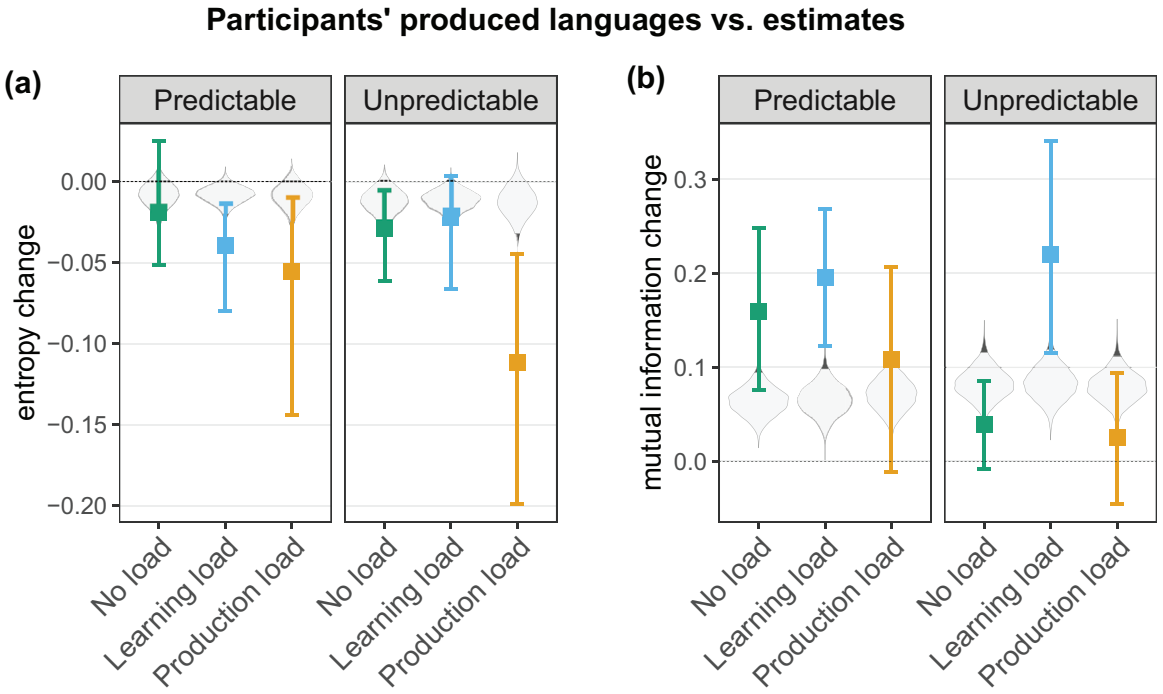

Fig. 4. Change in entropy (left) and mutual information (right) between the languages described by participants' estimates and the ones they produced, by condition. Points represent condition means; error bars represent bootstrapped 95% confidence intervals over the mean. Violins show the distribution of expected means under the null hypothesis of probability matching; regularization is indicated by means below the lower tail (entropy) or above the upper tail (MI) of these distributions. The same memory manipulation that drives regularization behavior during production also predicts how much more regular participants are in production than in their estimates in terms of entropy change. Participants produce a more deterministic pattern of conditioning than the one described by their estimates in the majority of conditions; however, in the UNPREDICTABLE/NO LOAD and UNPREDICTABLE/PRODUCTION LOAD conditions, learning effects account for all the increase in MI seen in production.

We also predicted that participants in all conditions would produce a more regular language than the one described by their estimates. This pattern is what was found by Ferdinand et al. (2019), who use it to argue that regularization is driven by production-side biases. In addition, we predicted that the same factors that drive regularization behavior during production should explain differences in regularity between participants' productions and their estimates. Taken together, we thus predicted that differences across conditions in the regularity of productions compared to input would be replicated when comparing productions to estimates.[13] In other words, when plotting the change in entropy and MI by condition, we would expect to see similar patterns for the production-input comparison and the production-estimate comparison.

Fig. 4a shows the difference in entropy. On this measure, participants were more regular in production than in their estimates in all conditions except PREDICTABLE/NO LOAD ($Z = -1.385$, $p = 0.90$) and UNPREDICTABLE/LEARNING LOAD ($Z = -1.579$, $p = .067$). In line with our prediction, the same memory manipulation that drives regularization behavior during production also predicts how much more regular participants are in production than in their estimates: permutation analysis confirms a main effect of memory load, with greater entropy

drop in PRODUCTION LOAD conditions than other memory load conditions ($Z = -2.527$, $p = .007$). As in the production-input comparison, permutation analysis shows no main effect of variation type ($Z = 0.796$, $p = .218$), and no interaction between variation type and memory load ($Z = 1.075$, $p = .150$).

Fig. 4b shows the difference in MI between participants' productions and their estimates. On this measure, participants were more regular in production than in their estimates in all conditions except UNPREDICTABLE/NO LOAD ($Z = -2.700$, $p = .997$) and UNPREDICTABLE/PRODUCTION LOAD ($Z = -3.394$, $p = .999$), suggesting that the increase in MI seen in participants' productions is accounted for by learning effects in these conditions. Unlike in the production-input comparison, permutation analysis shows no main effect of variation type ($Z = 1.635$, $p = .051$). However, as in the production-input comparison, permutation analysis reveals a main effect of memory load in the opposite direction than predicted: MI increases less in PRODUCTION LOAD conditions than others ($Z = -2.251$, $p = .011$). Exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions (collapsed) supports the notion that MI increases significantly more in LEARNING LOAD conditions than others ($Z = 3.102$, $p < .001$). Again, the interaction analysis we ran for the entropy data is clearly not warranted by the MI data since the main effect does not go in the predicted direction. We carried out a further exploratory analysis comparing LEARNING LOAD conditions to other memory load conditions (collapsed), but this analysis revealed no reliable interaction between variation type and memory load ($Z = 1.551$, $p = .060$).

To summarize, these results show, in line with our prediction, that reduction of overall variability is driven by memory limitations during language *production*. By contrast, lexical conditioning is boosted relative to the input almost across the board, and this tendency is even more pronounced when memory is taxed during *learning*.

Fig. 5 shows an example of one participant's behavior across the experiment. This participant was in the UNPREDICTABLE/LEARNING LOAD condition, so they were trained on a language with a 62.5/37.5 split between the two plural markers for every noun. Their estimates describe a very different language: one where four nouns *only* appear with the majority marker,[14] one noun *only* appears with the minority marker, and the remaining noun has a roughly 50/50 split between the two plurals. This language has entropy of 0.82 (compared to the input entropy of 0.95) and MI of 0.64 (compared to the input MI of 0). The language they produced was even more regular than their estimates in terms of lexical conditioning, with MI of 0.85, but almost identical to the input in terms of the overall frequency distribution of plural markers, with entropy of 0.94.

2.3. Discussion

In this experiment, we investigated whether working memory limitations during production drive regularization of both predictable (conditioned) and unpredictable (random) variation. In line with this hypothesis, we found evidence for a reduction in both types of variation when memory was taxed during production. As in previous research (e.g., Ferdinand et al., 2019; Hudson Kam & Newport, 2009; Saldana et al., 2021; Schwab et al., 2018), this effect was not driven by learners failing to accurately encode the overall frequency of different variants

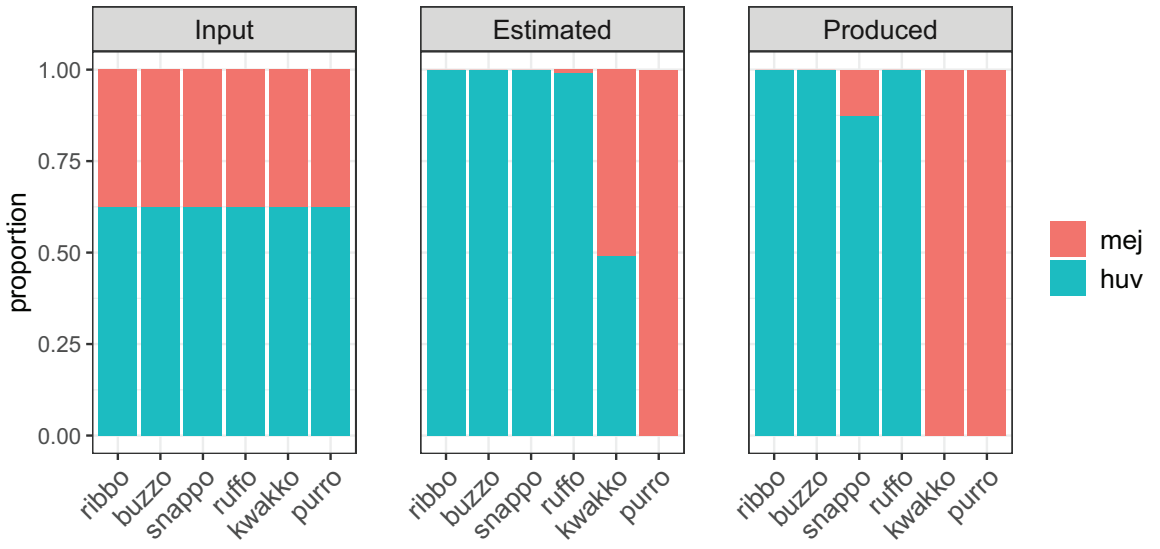

Fig. 5. Example language estimated (middle) and produced (right) by one participant in the UNPREDICTABLE/LEARNING LOAD condition, relative to the input (left). The participants' estimates indicate that they learned a pattern of lexical conditioning that was not present in the input; they then made this pattern even more deterministic in production.

in their input. Importantly though, our results do not support an exclusively production-side account of regularization. In particular, we found evidence for an increase in lexical conditioning during both learning and production. In other words, a bias to reduce variability by increasing conditioning affects both language users' inferences during learning and their (implicit) decisions during production.

Although we saw a reduction in overall variation when taxing memory during production (a drop in entropy), we also observed a different kind of regularization in this experiment: an increase in lexical conditioning. However, this effect was *not* driven by memory load during production and was, if anything, amplified by taxing memory during *learning*. This suggests that working memory limitations during language production can account for regularization at the system-wide level but not at the lexical level. In other words, language users might overproduce particular variants (relative to their frequency in the input) as a result of limitations on memory retrieval, but this is not the mechanism by which variation becomes lexically conditioned. This begs the question: What assumptions do we need to make about memory retrieval processes in order to explain this discrepancy? In other words, *how* do limitations on working memory during language production give rise to some properties of regularity but not others? We turn to this question in the next section.

3. A model of production-side regularization

Historically, computational work has sought to explain regularization as a function of *learning* biases (e.g., Culbertson et al., 2013; Perfors, 2012; Ramscar & Gitcho, 2007; Ramscar & Yarlett, 2007; Reali & Griffiths, 2009; Rische & Komarova, 2016). However, in our

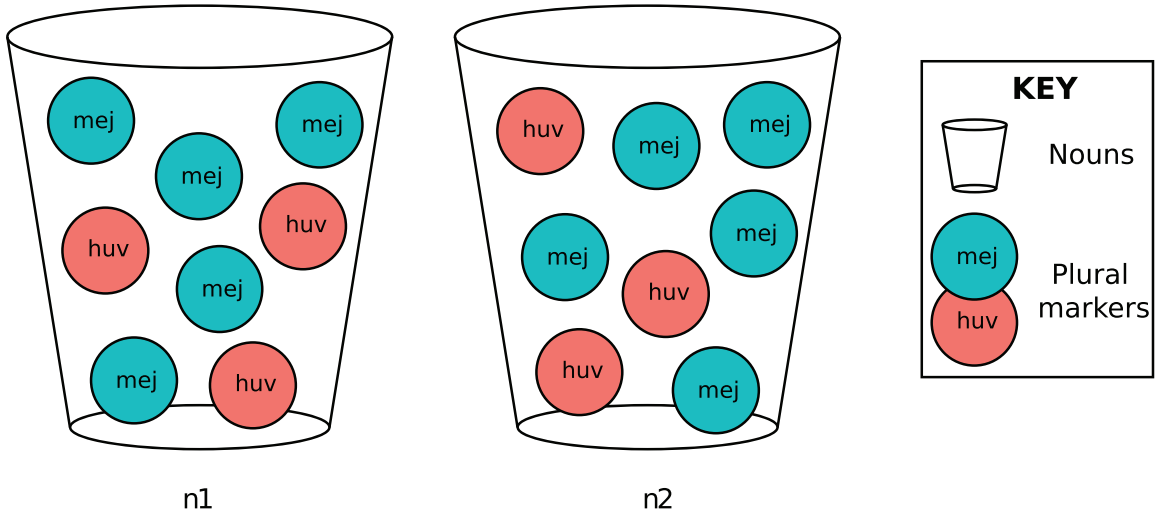

Fig. 6. An urn model conceptualization of nominal plural marking. Plural markers are represented as balls in urns (nouns). When agents encounter a noun n_i, they produce a plural marker by choosing a ball at random from the associated urn U_{n_i}. In this case, the agent would produce "mej" with a probability of 0.625 for either noun.

experiment, we found that working memory limitations operating during language *production* were a reliable predictor of regularization behavior. Furthermore, learning data (from the estimation task) did not reveal any prior bias for the kind of regularity we saw emerging in PRODUCTION LOAD conditions, that is, an overall loss of one variant in favor of another.

Previous work has suggested that the mechanism by which memory constraints result in regularization is overretrieval of a more accessible form (Goldberg & Ferreira, 2022; Hudson Kam and Chang, 2009; Marcus et al., 1992). More specifically, recent research (Hudson Kam, 2019; Schwab et al., 2018) has speculated that a kind of repetition priming might drive increased accessibility of forms that have been produced more recently. Here, we implement this mechanism in a simple "urn" model (Hintzman, 1986; Nosofsky, 1986; Spike, Stadler, Kirby, & Smith, 2017; Walsh, Möbius, Wade, & Schütze, 2010). We show that such a model can capture the entropy decrease in our experimental PRODUCTION LOAD conditions by means of a production process that causes one variant to be retrieved more than would be predicted by its frequency in the input.

3.1. Details of the model

Urn models represent the object of interest (here, plural markers) as balls in an urn or set of urns (here, nouns), where different variants correspond to different colored balls (Fig. 6). In the basic urn model, an agent draws a ball randomly from an urn and observes its color, places it back in the urn, and then repeats the selection process. Here, we model the memory load effect as a simple self-priming mechanism using a Pólya urn model (see Mahmoud, 2008, for an overview). In a Pólya urn model, k additional balls of the same color are added to the urn after each draw. In this way, the probability of producing a particular variant depends not only on that variant's frequency in the input but also on the frequency with which it has

already been produced; observed values become more likely to be observed again. In other words, variants that are produced more become even more accessible for retrieval in future trials than would be predicted by the input statistics alone. Note that this process does not inevitably favor the variant that had a higher frequency in the input: as long as an urn contains both variants, it is always possible that the lower frequency one will be chosen on the first trial and then boosted by the priming mechanism.

The population is a set of agents A who each learn a language L. Here, the language is a set of nouns $\{n_1, \ldots, n_6\} \in N$, each with an associated urn U_{n_i} containing plural markers from the set $\{p_1, p_2\} \in P$. Since participants in the real experiment did not always learn the input languages perfectly, we used the languages described by participants' estimates as the input to our agents. In this way, we can model the effect of production mechanisms *after* taking learning effects into account.

Each agent a completes 48 production trials—eight for each noun (as in the real experiment). On each trial, the agent encounters a random noun n_i and produces a plural marker for that noun by sampling the corresponding urn U_{n_i}. L is then updated according to the parameters described in the next section.

3.1.1. Parameters

In order to find a model that would provide the best fit to the experiment data, we consider all combinations of the following parameter settings for both PREDICTABLE and UNPREDICTABLE input languages. These parameters are intended to spell out the details of how self-priming through repeated production can give rise to regularization, and where this behavior comes from—both at an individual and population level.

Priming scope: One possibility is that priming is context-sensitive: the variant that was most recently produced for a given noun is more likely to be produced the next time *that noun* is encountered. Alternatively, priming could be context-agnostic: the variant that was produced on trial t_i is more likely to be produced on trial t_{i+1}, regardless of which nouns are encountered on those two trials. The *priming scope* parameter, therefore, has three possible values: within nouns, between nouns, or both.

Priming strength: Although we did observe regularization at a population level in our experimental PRODUCTION LOAD conditions, there was substantial variation in the extent to which individual participants showed this effect. We, therefore, wanted to allow agents in the model to differ systematically from each other in the same way. To do so, we randomly select a value of k for each agent: the number of additional balls they add to the relevant urns after each draw. Thus, the strength of the priming mechanism is a property of individuals, not a property of populations. We allow k to range between 0 and 8: at most, agents can add the same number of balls as were in the urn to start with, but it is possible for them to add none (and they can never take any away). Two parameters control the way k is selected.

First, we model the distribution of k in the population according to one of three distributions from the beta-binomial family: uniform ($\alpha = \beta = 1$), normal-like ($\alpha = \beta > 1$), or u-shaped ($\alpha < 1, \beta < 1$). These distributions capture different types of populations. In the uniform

distribution, all values of k are equally likely; in such a population, there is no concept of a "typical" agent. In the normal-like distribution, values around the mean are the most likely and extreme values (in either direction) are less likely. In the u-shaped distribution, extreme values are *more* likely. Specifically, we parameterize this distribution such that the most likely value is 0, the maximum value (given the range) is about half as likely, and values in the middle are considerably less likely. Concretely, approximately 90% of agents will use a value of k at one of the two extremes of the range.

Second, we consider all mean values of k in the set {1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0} for each distribution.[15] We use this value to set the upper bound on the range of allowable values.[16]

We sample k according to the following procedure:

$$k \leftarrow \begin{cases} \text{random.betabinom}\ (n = 2m,\ \alpha = 1,\ \beta = 1), & \text{if } d == \text{“uniform”} \\ \text{random.betabinom}\ (n = 2m,\ \alpha = 100,\ \beta = 100), & \text{if } d == \text{“normal-like”} \\ \text{random.betabinom}\ (n = 2m,\ \alpha = 0.05,\ \beta = 0.1), & \text{if } d == \text{“u-shaped”} \end{cases}$$

where m is the *mean priming strength* and d is the *population distribution*.

Forgetting: In the basic Pólya urn model, the number of balls increases at every time step when $k > 0$. We do not consider this situation here, since it would have the somewhat implausible effect that agents who are most affected by the self-priming mechanism (i.e., those with the most severely limited working memory) would also end up storing the largest number of data points in memory. An alternative model is one where the amount of data remains constant through the deletion of k balls from each urn for k that are added. We consider two deletion methods: either k balls are randomly removed from the urn, or deletion always targets the k oldest balls. The *forgetting* parameter, therefore, has two possible values: random or oldest. Importantly, forgetting never preferentially targets the low-frequency variant (a condition that was proposed to be essential in modeling of regularization during learning by Perfors, 2012).

3.1.2. Analysis

Each model is a unique combination of parameter settings. We ran 100 simulated experiments with each model, each consisting of the same number of agents in the PREDICTABLE and UNPREDICTABLE conditions as in the corresponding PRODUCTION LOAD conditions in the real experiment. For each experiment, we calculated the mean change in entropy and MI (relative to the input) by condition and obtained a 95% confidence interval around these means through bootstrapping. We then averaged over the 100 experiments. To determine which model provides the best fit to the experiment data, we compared these simulated means and confidence intervals to the corresponding means and confidence intervals of the PRODUCTION LOAD conditions in the real experiment. Each model received a divergence score, which captures the average absolute difference between the real and simulated means and confidence intervals across conditions; lower scores indicate that the data generated by that model are more similar to the real data.

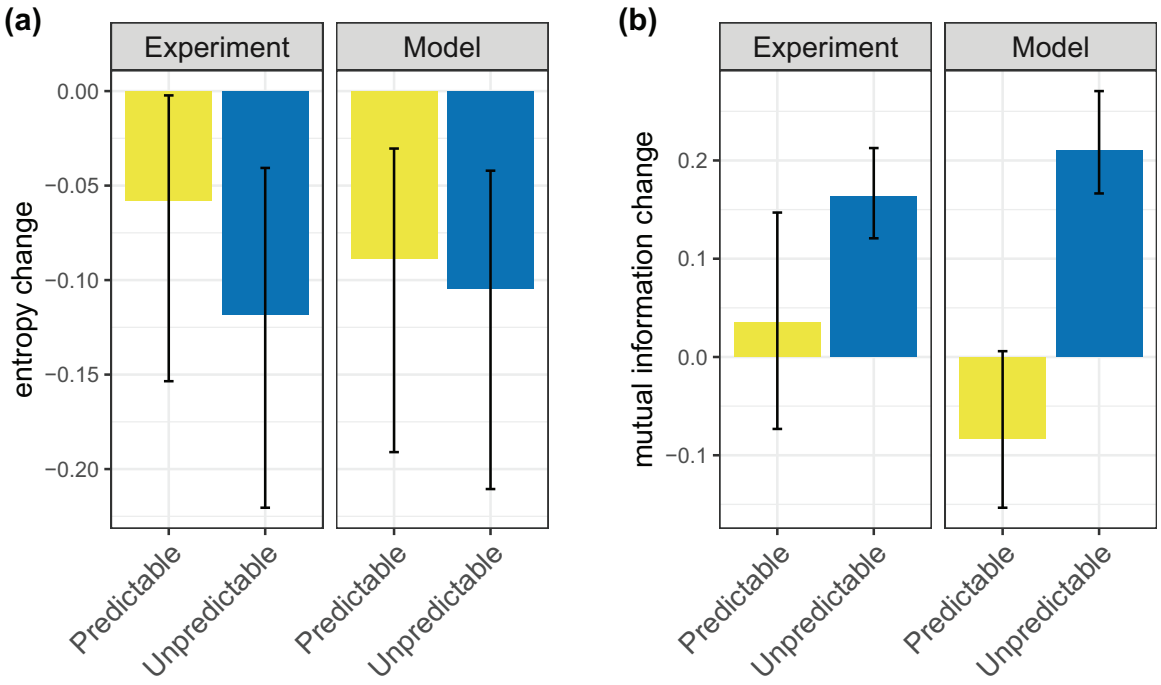

Fig. 7. Change in entropy (left) and mutual information (right) between input language and production output for participants in the experimental PRODUCTION LOAD conditions and agents in the best-fit model. Parameter settings were as follows: priming both within and between nouns, k (the priming strength parameter) drawn from a u-shaped distribution with mean 2.0, and random forgetting.

3.2. Model results

Fig. 7 shows the data generated by the model that provided the best fit to the experiment data overall. This model had priming both within and between nouns. The priming strength parameter k was drawn from a u-shaped distribution with median 2.0, that is, k could take any value in the set $\{0, 1, 2, 3, 4\}$, but extreme values were more likely. Balls were randomly selected for deletion after new ones were added. Further details of the performance of different parameter settings are available in Appendix B.

The inter-agent variation that arises by sampling k from some distribution on an agent-by-agent basis is a demonstrably key component of these models. Fig. 8 shows the entropy results for two models where all agents use the same value of k: either 1 (the lowest possible non-zero value) or 4 (the highest possible value in the distribution used by the best-fit model). When k is uniformly low, the model *under*-estimates both the mean decrease in entropy and the amount of variance around this mean (as indicated by narrower confidence intervals for the model than the experiment). When k is uniformly high, the model *over*-estimates the decrease in entropy for both conditions. These results provide further evidence that the data we observed in the experiment were generated by a population where individuals differ systematically in their sensitivity to the memory load manipulation. Specifically, the superior performance of the u-shaped distribution is suggestive of the nature of these individual differences: in our experiment at least, it seems likely that we were dealing with a population where most people were unaffected by the memory load manipulation, but those who were affected were extremely so.

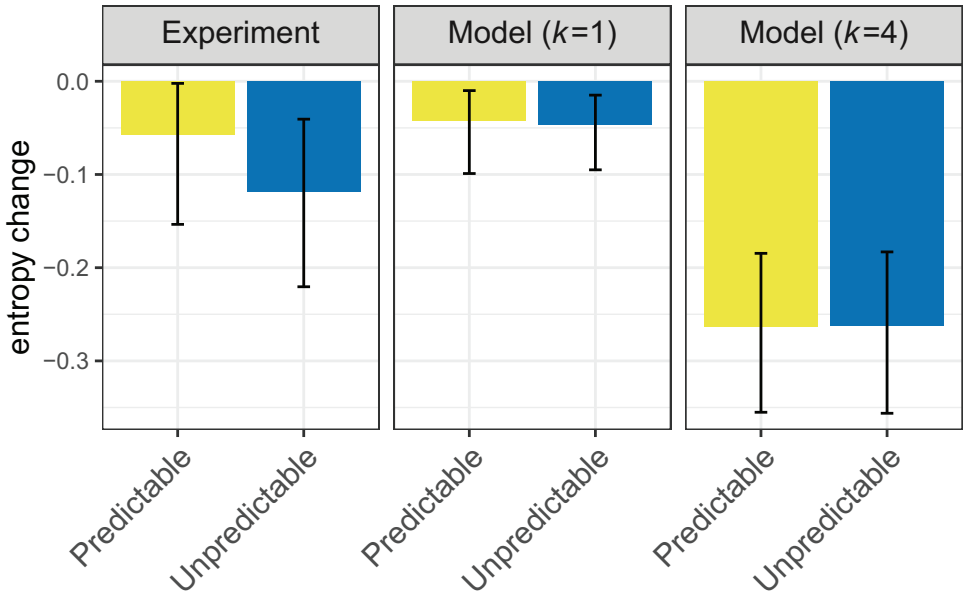

Fig. 8. Change in entropy between input language and production output for participants in the experimental PRODUCTION LOAD conditions and agents in two models with no inter-individual variation in priming strength. These models use the same settings as the best-fit model discussed above for the *priming scope* and *forgetting* parameters; the *population distribution* and *mean priming strength* parameters are not relevant when agents all use the same value of k (the priming strength parameter). When k is low, the model underestimates the true decrease in entropy. When k is high, the model *over*-estimates the decrease in entropy. These models demonstrate the importance of individual differences in priming strength for capturing the experiment results.

Finally, when agents sample from their input with no priming between trials (i.e., $k = 0$ for all agents), there is no significant drop in entropy: results mirror those of the experimental NO LOAD conditions (Fig. 9). This underlines the importance of a production-side mechanism for capturing the experiment results; imperfections in the learning process are not enough to explain the drop in entropy during production.

3.3. Discussion

With this model, we have shown that production mechanisms alone can give rise to levels of regularization comparable to those seen in our experiment, without the need for any prior bias against variability. Specifically, the mechanism implemented by our Pólya urn model can be thought of as a kind of self-priming: rather than agents sampling faithfully from the data they learned, the production process distorts the representation of that data that they draw on during production such that a recently produced variant becomes even more accessible for retrieval in future. Importantly, this distortion is frequency independent: none of our parameter settings involve preferential forgetting of irregular items or preferential retrieval of regular items (cf. Perfors, 2012, where such a model was argued to be the only way that regularization could arise from memory limitations during *learning*). Thus, the skew in the input itself provides the necessary conditions for regularization to occur under a neutral self-priming process.

In our experiment, we saw that the direction of travel was the same for both predictable and unpredictable variation—towards regularity. However, at least descriptively, unpredictable

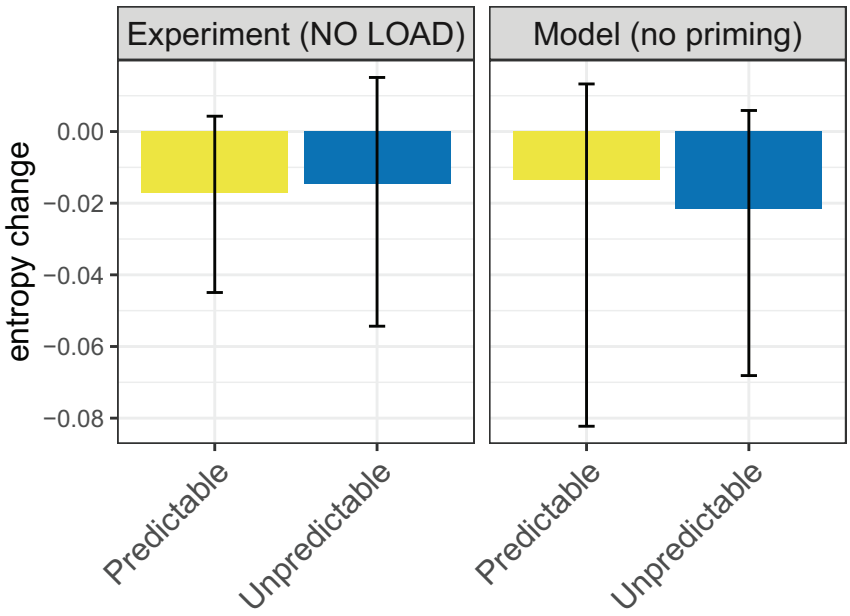

Fig. 9. Change in entropy between input language and production output for participants in the experimental NO LOAD conditions and agents in a model with no priming ($k = 0$ for all agents). When agents sample faithfully from their input, results mirror those of the experimental NO LOAD conditions, that is, no evidence of regularization.

languages tended to change more. One important aspect of these models is that the same parameter settings, applied to the two language types, generate a similar asymmetry. In other words, there is no need to posit different production-side biases targeting the different types of variation: the properties of the input—and differential learning of the two language types—naturally give rise to different amounts of regularization. However, it is true that our models generally perform better in the UNPREDICTABLE condition.

Finally, our results suggest that inter-individual variation in the strength of the priming mechanism is a key ingredient; when priming strength is uniform across the population, the models provide a poor fit to the experiment data. Furthermore, the best model of the population is one in which individuals differ from each other quite radically: most agents fall at one of the two extremes of priming strength, with very few in the middle.

Although our aim here was simply to provide a model that could account for our experimental data, future work could look to apply the mechanism we suggest to other aspects of natural language production that might be relevant to regularization. For example, our experiment does not involve generalization to novel nouns, but this is certainly a task that can increase the tendency to regularization (e.g., Wonnacott & Newport, 2005). Our model could be extended to account for this: the use of a given variant with one noun would prime that variant for all future nouns, whether or not those nouns have been seen before. Furthermore, in both our experiment and model it was not possible to innovate new forms, which removes one potential source of *irregularization*. This could be accounted for in the model through the addition of an error rate parameter which allows for occasional distortions of the sampling process, for example, the addition of an unattested ball to an urn.[17]

4. General discussion

In this study, we have added to a growing body of evidence showing that, even when linguistic variation is accurately learned, it is not always accurately reproduced (e.g., Austin et al., 2022; Ferdinand et al., 2019; Hudson Kam and Chang, 2009; Hudson Kam & Newport, 2009; Saldana et al., 2021; Schwab et al., 2018). Specifically, we have shown that constraints on language production arising from memory limitations can result in the loss of both predictable and unpredictable variation. However, we also found evidence that regularization is not exclusively an effect of production: the process by which random variation becomes conditioned is better explained by learning biases. Humans are powerful statistical learners across many domains, extracting even subtle regularities after very little exposure (see Saffran & Kirkham, 2018, and Sherman, Graves, & Turk-Browne, 2020, for reviews). However, people generally have poor perception of randomness and are quick to infer that random sequences are actually structured (Bar-Hillel & Wagenaar, 1991; Gaissmaier & Schooler, 2008; Hyman & Jenkin, 1956; Wolford, Newman, Miller, & Wig, 2004). Our results suggest that this bias generalizes to language acquisition, causing learners to identify and internalize regularities even when none existed in their input (Samara et al., 2017; Smith & Wonnacott, 2010).

4.1. Memory limitations: Learning or production effects?

In this study, we simulated the memory pressures inherent to language learning and production through a concurrent load task (Hudson Kam, 2019; Perfors, 2012). Of course, language users are not habitually asked to memorize and recall digit sequences during conversation, so this is a somewhat artificial view of working memory's role in language learning and use. Nonetheless, if disrupting working memory during particular linguistic tasks has behavioral consequences, we can infer that memory is a relevant constraint on those tasks generally.

Both our experimental and computational results suggest that memory limitations during language production can account for regularization at the global level (i.e., an overall increase in the frequency of one variant to the exclusion of others) but are not a particularly good predictor of regularization at the lexical level (i.e., the introduction of lexical conditioning). This discrepancy makes sense considering the mechanism that we are proposing for the production effect, whereby variants with a higher frequency (in either the observed data or in the output) become ever more accessible, and therefore ever more likely to be retrieved for production (Goldberg & Ferreira, 2022; Hudson Kam and Chang, 2009; Schwab et al., 2018). Introducing lexical conditioning, on the other hand, requires participants to boost the high-frequency variant for some nouns and the low-frequency variant for others, a process that cannot be easily explained under a memory retrieval account.

In fact, our exploratory analysis suggests that memory limitations during *learning* may have a role to play in explaining the evolution of predictable patterns of variation. At first glance, this result appears to dovetail with some earlier work in the "Less is More" tradition (Newport, 1988; Newport, 1990). For example, simulated agents and recurrent neural networks have been shown to learn linguistic regularities better when they begin with some kind of memory limitation—or input filter—and gradually mature (Elman, 1993; Goldowsky &

Newport, 1993). It has been suggested that, by limiting the size of the sample from which learners can draw inferences, these input filter mechanisms enhance the detection of meaningful relationships (Kareev, 1995; Kareev, Lieberman & Lev, 1997). However, more recent reanalysis (Brooks & Kempe, 2019; Rohde & Plaut, 2003; Rohde & Plaut, 1999) calls many of these findings into question. Specifically, Rohde and Plaut (2003) point out that although filtering mechanisms sometimes isolate the correct regularities, they just as often destroy important parts of the data and identify spurious regularities instead. Indeed, this is exactly what we see here: learners subject to the LEARNING LOAD manipulation are *less* successful at faithfully reproducing the language they were exposed to because they are detecting patterns that did not exist in their input.

Overall, our results lend support to the idea that regularization arises from memory constraints during language production but also suggest that this is not the whole story. If we consider regularization as the process by which language becomes more systematic and predictable—whether by reducing the number of variants in a system, or by specializing different variants for different contexts—then it appears that memory limitations are also doing something important during learning.

4.2. Revisiting the relationship between predictable and unpredictable variation

One of the key aims of this study was to investigate whether linguistic variation is a single phenomenon, with predictable and unpredictable variation constituting two points on the same spectrum. The implication of such a characterization is that the same kinds of biases should act on both types of variation. In other words, the same mechanisms that have been shown to result in regularization of unpredictable variation should also target predictable variation. Our results are consistent with this account when it comes to the effect of memory limitations during language *production*.

However, our analysis also suggests that truly random variation is subject to distortion during the learning process in a way that conditioned variation is not—even when that conditioning is only probabilistic. In other words, even though learning biases *could* theoretically have obscured the small amount of noise in our predictable languages, in fact participants' estimates show that they were very aware of this noise and did not believe that they had been exposed to a deterministic pattern of conditioning. Therefore, it appears that there may be something special about unpredictable variation when it comes to learning. Specifically, our results suggest that language learning is biased in favor of predictable dependencies between elements in a system to the extent that even random systems will be analyzed as containing such patterns. Future research could investigate how these learning biases play out across the spectrum of variation; for example, a less deterministic version of our predictable language might be subject to more distortion in learning.[18]

We observed two related but distinct biases in this study: a bias against variability of all kinds (driven by production) and a bias against unpredictability (driven by both learning and production). However, the second of these biases appears to be stronger: In both learning and production, we saw much bigger changes in MI than in entropy. A question for future work is how the relative strength of these biases interacts with the size of the system: with only two

variants, as in our design, it is presumably not difficult to maintain both. Expanding the language may heighten the pressure to regularize by losing some variants altogether, rather than just by introducing conditioning (although see Hudson Kam & Newport, 2009). Our forced-choice production task also minimized the kind of retrieval difficulties that we might expect to result in increased use of one variant to the exclusion of others since participants were cued to remember that there was another option even if they would have spontaneously favored a single variant. We would expect a different kind of production task—with participants required to free-type or orally produce their descriptions—to give rise both to a greater drop in entropy overall (Hudson Kam and Chang, 2009), and to a stronger effect of the interference task, especially if the language was more complex.

Overall, our results suggest that pragmatic factors alone cannot fully explain regularization and that working memory limitations offer a plausible *cognitive* explanation for this phenomenon. Specifically, we found that increased cognitive load during language production gave rise to increased regularization of both predictable and unpredictable variation—in the absence of any communication between participants or differences in pragmatic framing of the task. Furthermore, a pragmatic account would not predict any regularization during learning, since the mechanisms implicated in such accounts only come into play during production. However, our results clearly show that when participants produce a more predictable language than the one they were exposed to, this is at least partly because they have failed to accurately learn the randomness in their input.

4.3. Why do languages have variation at all?

Our results suggest that biases arising from memory limitations broadly disfavor linguistic variation, even when that variation is predictable. From the perspective of language evolution, one might, therefore, wonder why variation is so pervasive in natural languages. As with any cognitive bias shaping language, the explanation for this is likely a combination of the fact that these biases are weak (i.e., defeasible) and compete with other pressures shaping language. Most obviously, patterns of linguistic usage are influenced by the social contexts in which they are found: there is ample evidence to suggest that speakers use variation as a marker of social identity (see Chambers & Schilling, 2018, for an overview). Furthermore, some types of variation may be preferred because of cognitive biases pertaining to specific linguistic or semantic categories (e.g., Christensen, Fusaroli, & Tylén, 2016; Holtz, Kirby, & Culbertson, 2022; Motamedi, Wolters, Naegeli, Kirby, & Schouwstra, 2022; Napoli & Sutton-Spence, 2014; Schouwstra & De Swart, 2014).

Individual differences in the strength of the regularization bias may also help to explain how variation can persist in natural language. Our experimental data certainly suggest that memory load does not lead to regularization across the board. In particular, a wide range of behaviors were represented in our PRODUCTION LOAD conditions. Many participants in these conditions seemed not to be hindered at all by the interference task, producing languages with near-zero entropy change compared to the input.[19] Some appeared to be moderately affected, maintaining some but not all of the variation that was present in the input. And a small handful were severely disrupted, producing only one variant in testing. Similarly, our

best-fit computational model was one in which the majority of agents actually had no propensity towards regularization, but those who did tended to reduce variation quite substantially. In terms of diachronic change in natural language then, if only some individuals have very strong biases against variation, we should perhaps not expect that variation to be lost either quickly or completely.

Finally, although this was not relevant in our study, systems of conditioned variation may persist due to frequency-dependent patterns in regularity. Irregular forms tend to be highly frequent, presumably making it easier to learn and retrieve the correct form (Cuskley et al., 2014; Wu, Cotterell, & O'Donnell, 2019). Furthermore, learners are sensitive to the frequency of specific exemplars (e.g., the frequency of the word *went*) as well as the frequency of morphological types (e.g., the frequency of the *-ed* past tense marker), so it is not necessarily the case that the "regular" variant is the most easily retrieved in all contexts (Arnon, 2015; Arnon & Snider, 2010). Indeed, usage-based models (e.g., Bybee, 2006; Bybee, 2002; Hay, 2001; Langacker, 1988) argue that the easiest variant to access in any given context is simply the one that has been experienced most often in that context. In such models, linguistic data form memory representations whereby items that are experienced frequently together start to form a unit; these units then come to be processed and retrieved holistically and thus become resistant to restructuring (Bybee, 1985; Bybee & Thompson, 1997). There is also growing recognition that learners actually *start out* with such holistic units in some cases, especially for high-frequency items (Arnon & Clark, 2011; Chevrot, Dugua, & Fayol, 2008; Christiansen & Arnon, 2017; Havron & Arnon, 2021; Lieven, Pine, & Baldwin, 1997; Pine & Lieven, 1997; Siegelman & Arnon, 2015). In this case, lexically conditioned variation may persist because highly frequent irregular items never get segmented, and thus, when producing these items, there is no process of retrieving individual morphemes during which an alternative form could be retrieved (cf. Pinker & Ullman, 2002). Therefore, while regularization might arise when a high-frequency type is extended to a less familiar context (Harmon & Kapatsinski, 2017; Koranda, Zettersten, & MacDonald, 2018; Wonnacott, 2011), high-frequency irregular items are likely to be evolutionarily stable. All nouns in our design were equally frequent, so our results do not speak to any potential relationship between frequency and memory limitations in driving regularization. However, the paradigm we present here could certainly be used to test this hypothesis.

5. Conclusion

We have provided evidence that cognitive biases leading to regularization target both unpredictable and predictable variation. Our findings support the idea that regularization is particularly strong during production and is driven at least in part by memory limitations. However, our results also suggest that this is not the whole story; while over-retrieval of a more accessible variant during language production may act to reduce overall variability, *unpredictability* appears to decrease more as a result of inferences formed during learning. Overall, this study lends support to the notion that cognitive constraints in individuals can give rise to particular structures in languages. Specifically, we argue that—all things

equal—regularities that allow languages to pass more easily through the bottleneck imposed by working memory limitations will tend to accumulate as languages evolve, leading to the appearance of typological universals.

Acknowledgments

This project was supported by funding from the Economic and Social Research Council (grant ref. ES/P000681/1, awarded to AK) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643, awarded to JC). We are grateful to Elizabeth Pankratz for drawing the stimuli, and to members of the Centre for Language Evolution for helpful discussion. Thank you to our reviewers, Carla Hudson Kam and two anonymous reviewers, for valuable input.

Open Research Badges


This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/9e27b

Notes

1 Either by producing an invalid label type (*noun* + *noun*, *marker* + *marker,* or *marker* + *noun*; the only valid label type was *noun* + *marker*), or by producing an incorrect noun.
2 Defined as clicking buttons in the same left-right position on more than 90% of trials.
3 Due to a technical error, the order of presentation was not fully randomized in this phase. Instead, all participants saw eight passes through the stimuli set in the same randomized order each time. We have no reason to expect that this would have affected participant behavior.
4 This potentially reduces the strength of regularization behavior compared to a free production task, but it is still a task where regularization can be observed with the right analysis techniques, for example, Ferdinand et al. (2019)
5 For example, due to primacy or recency effects (Ferdinand et al., 2019).
6 $H(V) - H(V|C)$.
7 Note that, although participants could in principle produce a language with MI of 1, this is not what we would see if they simply produced a deterministic version of the conditioning pattern in their input (i.e., four nouns with one marker and two with the other); such a language would score 0.92.
8 The pattern of results under this analysis is identical to the one obtained from our preregistered linear models.
9 Note that we do not draw any inferences from significantly positive z-scores for entropy change or significantly negative z-scores for MI change: none of our predictions are

about an increase in variability, so our focus is simply on whether there is or is not evidence for regularization. In other words, these are all one-tailed tests.

10 Shuffling in this way will, on average, give the same mean in each condition, so the resulting distribution will be normal and centered around 0, that is, no difference between conditions.

11 This resulted in the exclusion of 322 (out of 8,433) trials. Of these, the word occupying the noun slot was an incorrect noun on 267 trials, of which the label was a valid *noun + plural* form on 168 trials; on the remaining 99 trials, the word occupying the plural slot was another noun, suggesting that the participant had tried to correct their mistake with their second click (as indicated in some debrief questionnaires). Of the remaining 55 trials, there were 34 cases where the noun was correct but the label was invalid because the noun had been duplicated. In 11 cases, the noun was correct but the words were in the wrong order (i.e., the label was of the form *plural + noun*).

12 Note that there was more scope for MI to increase in UNPREDICTABLE conditions because the starting point (0) was lower for these languages than in PREDICTABLE conditions (0.41).

13 In this case, we compare the real by-condition means to the mean of a corresponding simulated condition where participants probability match their *estimates* (rather than the input).

14 Rounding up for "ruffo": this slider was set to 99%.

15 A mean greater than 4 would allow k to take values outside of the defined range.

16 Strictly speaking, this parameter controls the *median* of the u-shaped distribution rather than the mean, since the distribution is asymmetric.

17 Thank you to an anonymous reviewer for these interesting suggestions.

18 Thank you to an anonymous reviewer for this suggestion.

19 Although we would expect to see fewer participants in this category if the production task itself was more taxing, that is, free production rather than forced-choice.

References

Aitchison, J. (1996, July 20). Small steps or large leaps? Undergeneralization and overgeneralization in Creole acquisition. In H. Wekker (Ed.), *Creole languages and language acquisition* (pp. 9–32). Berlin, Germany: Mouton De Gruyter.

Arnon, I. (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, *42*(2), 274–277.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth - children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, *18*(3), 249–277.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation - Advances in Research and Theory*, *8*(100), 47–89.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bickerton, D. (1981). *Roots of language*. Berlin, Germany: Language Science Press.

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, *29*(5), 591–610.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the less-is-more hypothesis. *Language Learning*, *69*, 13–41.

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2022). Semantic cues in language learning: An artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531.

Bybee, J. (1985). *Morphology: a study of the relation between meaning and form*. Amsterdam, The Netherlands: John Benjamins.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. *Annual Meeting of the Berkeley Linguistics Society*, *23*(1), 378–388.

Carroll, R., Svare, R., & Salmons, J. C. (2013). Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics*, *2*(2), 153–172.

Chambers, J. K., & Schilling, N. (2018). *The handbook of language variation and change*. Hoboken, NJ: John Wiley & Sons.

Chevrot, J.-P., Dugua, C., & Fayol, M. (2008). Liaison acquisition, word segmentation and construction in French: A usage-based account. *Journal of Child Language*, *36*(3), 557–596.

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268–293.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, *5*(3), 392–424.

Culbertson, J., & Wilson, C. (2013). Artificial grammar learning of shape-based noun classification. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*, 2118–2123.

Cuskley, C., Castellano, C., Colaiori, F., Loreto, V., Pugliese, M., & Tria, F. (2017). The regularity game: Investigating linguistic rule dynamics in a population of interacting agents. *Cognition*, *159*, 25–32.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of English. *PLoS ONE*, *9*(8), e102882.

DeGraff, M. (1999). *Language creation and language change: Creolization, diachrony, and development*. Cambridge, MA: MIT Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*, 37–64.

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Gardner, R. A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology*, *70*(2), 174–185.

Givón, T. (1985). Function, structure and language acquisition. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 1005–1028). Mahwah, NJ: Lawrence Erlbaum Associates.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four… or is it two? *Memory*, *12*(6), 732–747.

Goldberg, A., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. *The Proceedings of the 24th Annual Child Language Research Forum* (pp. 124–138). Stanford, CA: Stanford Linguistics Association by the Center for the Study of Language and Information.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44.

Havron, N., & Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, *48*(2), 244–260.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(6), 1041–1070.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Holtz, A., Kirby, S., & Culbertson, J. (2022). The influence of category-specific and system-wide preferences on cross-linguistic word order patterns. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *44*, pp. 1011–1018). Austin, TX: Cognitive Science Society.

Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language*, *91*, 906–937.

Hudson Kam, C. L. (2019). Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Language Learning and Development*, *15*(4), 317–337.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3), 815–821.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Hyman, R., & Jenkin, Noel. S. (1956). Involvement and set as determinants of behavioral stereotypy. *Psychological Reports*, *2*(3), 131–146.

Johnson, J. S., Shenkman, K. D., Newport, E. L., & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, *35*(3), 335–352.

Kamps, C., Ferdinand, V., & Kirby, S. (2014). The origins of regularity in language: Why coordination matters. In Cartmill, E. A., Roberts, S., Lyn, H., & Cornish, H., (Eds.), *The evolution of language: Proceedings of the 10th International Conference* (pp. 457–458). Singapore: World Scientific

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, *126*, 278–287.

Karmiloff-Smith, A. (1981). *A functional approach to child language: a study of determiners and reference*. Cambridge, England: Cambridge University Press.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford, England: Oxford University Press.

Koranda, M., Zettersten, M., & MacDonald, M. C. (2018). Word frequency can affect what you choose to say. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40, pp. 629–634). Madison, WI: Cognitive Science Society.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 127). Amsterdam. The Netherlands: John Benjamins

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.

Mahmoud, H. (2008). *Polya urn models*. Boca Raton, FL: CRC Press.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–178.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Motamedi, Y., Wolters, L., Naegeli, D., Kirby, S., & Schouwstra, M. (2022). From improvisation to learning: How naturalness and systematicity shape language evolution. *Cognition*, *228*, 105206.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, *5*, 376.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American sign language. *Language Sciences*, *10*(1), 147–172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(4), 486–506.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*(2), 138–155.
Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.
Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.
Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, *18*(3), 571–590.
R Core Team. (2022). R: A language and environment for statistical computing. Vienna, Austria: R Project for Statistical Computing.
Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*(7), 274–279.
Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927–960.
Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.
Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1–30.
Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.
Rohde, D. L. T., & Plaut, D. C. (2003). Less is less inlanguage acquisition. In P. Quinlin (Ed.), *Connectionist modelling of cognitive development* (pp. 160–200). East Sussex, England: Psychology Press.
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203.
Saldana, C., Smith, K., Kirby, S., & Culbertson, J. (2021). Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order. *Language Learning and Development*, *17*(2), 158–188.
Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.
Schouwstra, M., & De Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436.
Schwab, J. F., Lew-Williams, C., & Goldberg, A. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language*, *45*(5), 1054–1072.
Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, *12*(4), 323–328.
Senghas, A., Coppola, M., Newport, E. L., & Supalla, T. (1997). Argument structure in Nicaraguan sign language: The emergence of grammatical devices. In Hughes, E., Hughes, M., & Greenhill, A., (Eds.), *Proceedings of the Boston University Conference on Language Development* (Vol. *21*, pp. 550–561)
Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.
Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20.
Siegel, J. (2007). Recent evidence against the language bioprogram hypothesis. *Studies in Language*, *31*(1), 51–88.
Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75.
Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive Psychology*, *49*(4), 370–407.
Smith, K., Ashton, C., & Sims-Williams, H. (2023). The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 851–857.
Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.
Smith, K. H. (1969). Learning Co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 319–321.
Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.
Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science*, *34*(4), 537–582.
Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, *58*(4), 221.
Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1–14.
Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In Brugos, A., Clark-Cotton, M., & Ha, S. (Eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*, (pp. 1–11). Somerville, MA: Cascadilla Press.
Wu, S., Cotterell, R., & O'Donnell, T. (2019). Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5117–5126). Stroudsburg, PA: Association for Computational Linguistics.

Appendix A: Individual-level experimental data

All plots in this appendix show individual participants as colored points and condition means as black points. Error bars represented bootstrapped 95% confidence intervals over the mean.

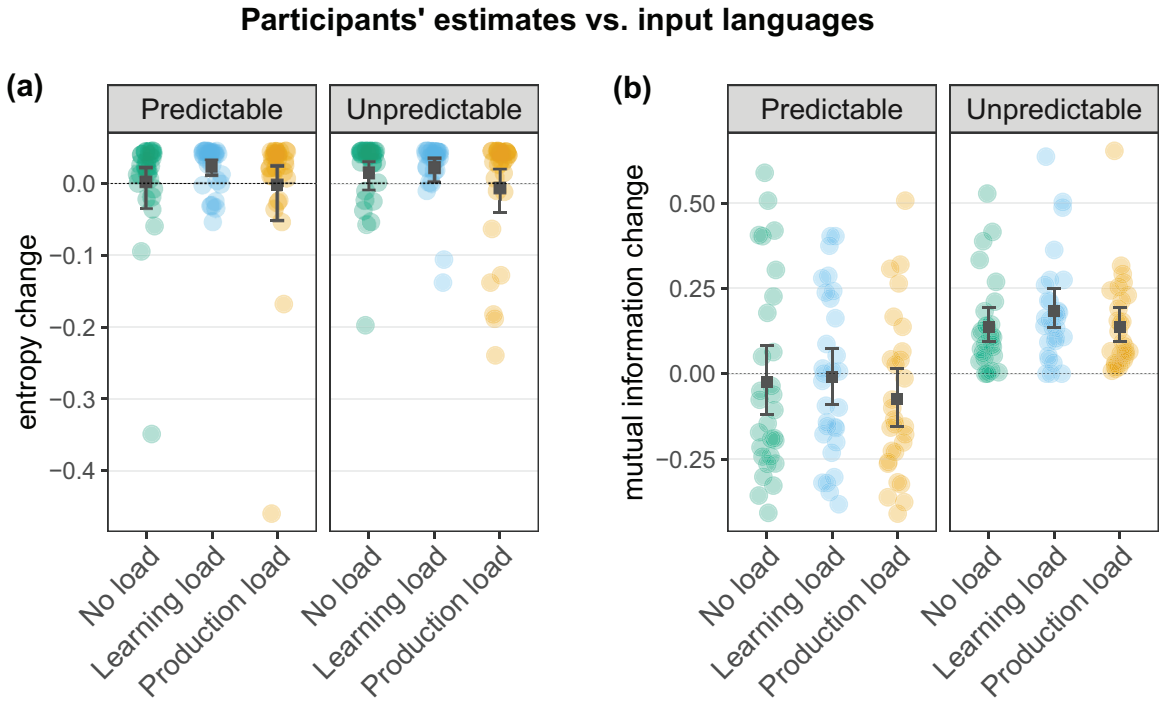

Fig. A.1. Change in entropy (left) and MI (right) between the languages participants were trained on and the ones described by their estimates, by condition.

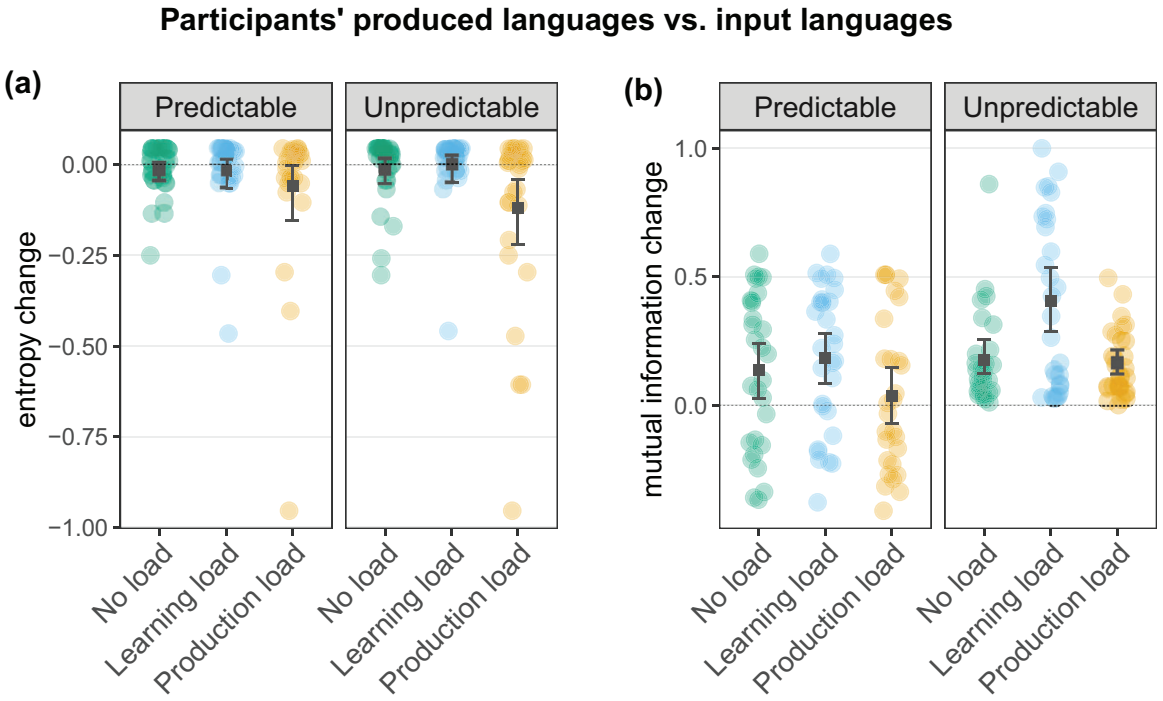

Fig. A.2. Change in entropy (left) and MI (right) between the languages participants were trained on and the ones they produced, by condition.

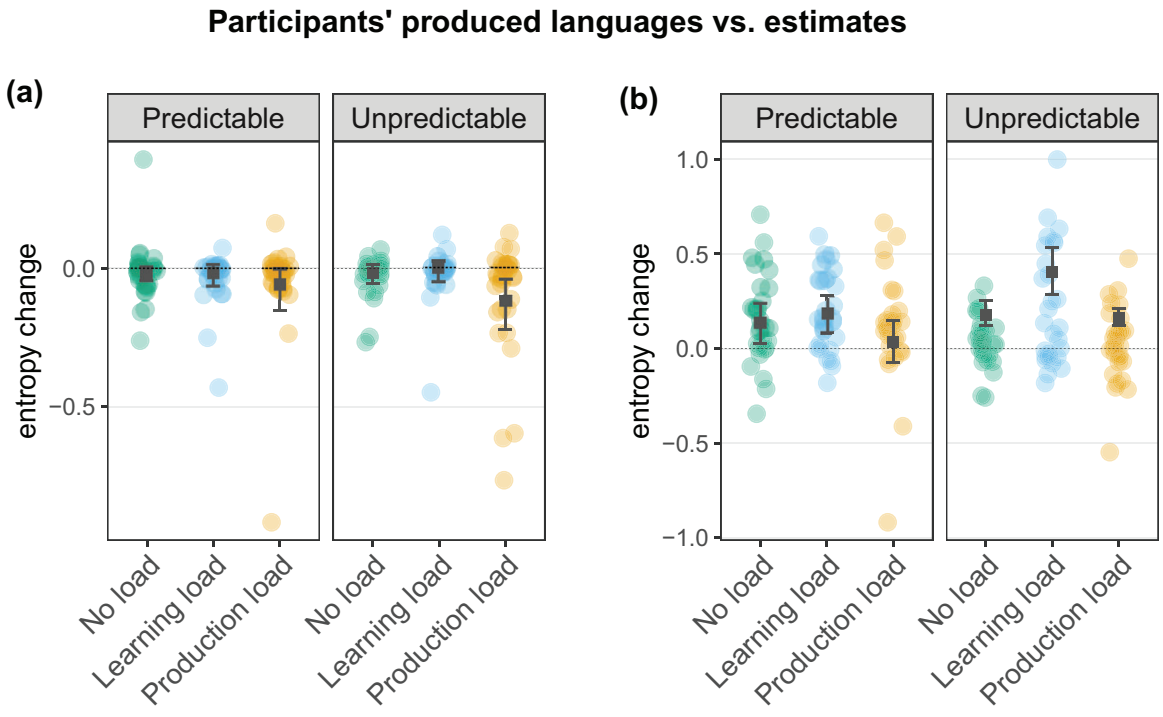

Fig. A.3. Change in entropy (left) and MI (right) between the languages described by participants' estimates and the ones they produced, by condition.

Appendix B: Additional model analysis

In general, all computational models were closer to the real data on entropy than MI, meaning that they were capturing the overall loss of variation better than the increase in lexical conditioning. Different settings of the *priming scope* parameter in particular generated very different results for the two measures, as shown in Table B.1. On average, "within-nouns" models performed considerably better than others on entropy. However, these models dramatically overestimated the change in MI relative to the experiment, since priming only within nouns leads to a very high likelihood of lexical conditioning (i.e., an increase in MI). In fact, the single best-fit model to the entropy data (divergence = 0.046) provided one of the *worst* fits to the MI data (divergence = 1.103), meaning that it was impossible to select a single model that could capture both effects in the experiment. "Between-nouns" models exhibited the opposite problem: while some models provided a reasonable fit for the UNPREDICTABLE condition, MI always *decreased* more in the PREDICTABLE condition than in the real experiment. Although "between-nouns" models had the lowest average divergence score overall, the single best-fitting model used the "within-and-between" setting. Moreover, these models had the most similar performance between entropy and MI.

The performance of different settings for the *mean priming strength* and *forgetting* parameters depended heavily on *priming scope* and varied between measures. For entropy, there was a negative correlation between *mean priming strength* and average divergence scores for "within-nouns" models and a positive correlation for others. In other words, higher means provided a better fit for "within-nouns" models, while lower means performed better for "between-nouns" and "within and between" models. Similarly, models with "oldest" *forgetting* performed marginally better than "random" models when priming was only within nouns, but considerably worse when priming was between nouns or both within and between. Overall, averaging over different settings of the *priming scope* parameter, higher means (Table B.2), and "oldest" forgetting (Table B.3) always provided a worse fit.

In terms of the *population distribution* parameter, there was relatively little difference between uniform and normal-like models (especially on entropy), but u-shaped models considerably out-performed both across the board (Table B.4). In fact, the top 10 best-fitting models overall all used the u-shaped distribution.

Table B.1
Divergence scores for different settings of the *priming scope* parameter

Priming Scope	Entropy	MI	Overall
Within nouns	0.122	0.618	0.370
Between nouns	0.298	0.374	0.336
Within and between	0.411	0.334	0.372

Abbreviation: MI, mutual information.

Table B.2
Divergence scores for different settings of the *mean priming strength* parameter

Mean priming strength	Entropy	MI	Overall
1.0	0.130	0.310	0.220
1.5	0.156	0.361	0.258
2.0	0.205	0.408	0.307
2.5	0.264	0.447	0.356
3.0	0.325	0.488	0.407
3.5	0.395	0.524	0.459
4.0	0.462	0.555	0.508

Abbreviation: MI, mutual information.

Table B.3
Divergence scores for different settings of the *forgetting* parameter

Forgetting	Entropy	MI	Overall
Random	0.149	0.376	0.262
Oldest	0.405	0.508	0.456

Abbreviation: MI, mutual information.

Table B.4
Divergence scores for different settings of the *population distribution* parameter

Population distribution	Entropy	MI	Overall
Uniform	0.306	0.476	0.391
Normal-like	0.306	0.537	0.421
U-shaped	0.219	0.313	0.266

Abbreviation: MI, mutual information.

Chapter 3

Task effects in morphological rule learning

> Found it quite easy, might work as a shorter study.
>
> *Anonymous Prolific participant (on an experiment with a 58% exclusion rate)*

Author contributions

This chapter was co-authored with a fellow PhD student, Elizabeth Pankratz, and a version (subject to different corrections) also appears in her thesis (Pankratz 2025). It has not been published in any other venue. Elizabeth and I jointly developed the research question, hypotheses and experimental design. Our supervisors, Jennifer Culbertson and Simon Kirby, provided advice on the experimental design during piloting, and suggested the follow-up with L1 German speakers. I created the experiment software, collected the data, and wrote the first draft of the methodology sections. Elizabeth designed the stimuli, analysed the data, created the visualisations, and wrote the first draft of the rest of the chapter. Elizabeth and I worked together to revise the draft and agree a final version of the chapter. Our supervisors both gave comments on draft versions.

Open materials

All materials, code and data used for this chapter are freely available at https://osf.io/zv4p6/.

3.1 Introduction

Learners of a new language will often discover that no matter how many target-language books, films, or podcasts they absorb, their language skills do not truly blossom until they have practised producing the language themselves.

Language production benefits both infant learners of their first language as well as adult learners of further languages. In L1 acquisition, children who use their target language more frequently show stronger expressive abilities in that language throughout development, independent of their level of comprehension (Bohman et al. 2010; Donnelly & Kidd 2021; Ribot et al. 2018). And in L2 acquisition, production tasks have been shown to improve how L1 Mandarin users learn English relative clauses (Izumi 2002) and how people with diverse L1s learn German grammatical gender (Keppenne et al. 2021); the way production tasks benefit L2 acquisition has been influentially referred to as the Output Hypothesis (Swain 2005). Artificial language learning studies also illustrate that production practice helps adults both to learn rules (Hopman & MacDonald 2018) and to generalise them (Hopman 2022).

A separate strand of research has shown that adult learners don't acquire all kinds of rules equally well. Particularly troublesome are morphological rules; adult learners' difficulty with both nominal and verbal inflectional morphology has been well documented (see e.g. Bentz & Winter 2013; Clahsen et al. 2010; DeKeyser 2005; Ellis 2022; Holmes & Dejean De La Bâtie 1999; Kenanidis et al. 2023; MacWhinney 2018; Parodi et al. 2004; Rogers 1987; Sagarra & Ellis 2013). Case marking poses a particular challenge: L2 learners of German and Turkish struggle to learn the case-marking morphology, even if their L1 also has case (Jordens et al. 1989; Papadopoulou et al. 2011). In contrast, rules that apply to larger chunks, such as words and phrases, seem more accessible (MacWhinney 2018; Sagarra & Ellis 2013). For example, when learning noun classification systems, adults tend to rely more on class membership cues that do not require them to segment below word level (i.e. determiners) compared to sub-word cues that do require segmentation (i.e. suffixes; Keogh and Lupyan 2024). And a recent study involving an artificial language that had both word order and case marking cues to thematic role showed that adult participants (L1 English and L1 German) relied more strongly on the word order cues, and that they were better able to

detect violations in word order than in case marking (Kenanidis et al. 2023); for further studies demonstrating adults' preference for word order rules over case marking, see Grey et al. (2014) and Rebuschat et al. (2021). Typological evidence also suggests that adults may prefer word-level rules: languages with more adult L2 learners tend to be morphologically simpler (Bentz & Winter 2013; Lupyan & Dale 2010).

In this study, we ask whether practising a new language with a more production-like task can help adult learners to acquire a hard-to-learn morphological rule that requires words to be segmented, moving beyond an easier word-level rule that requires no segmentation. This question builds on intriguing results from Hopman and MacDonald (2018). In their artificial language learning experiment, participants who did a production task seem to have acquired morphological rules better than word-level ones. Specifically, those participants appear to be more sensitive to errors in suffixing than errors in word order (see their Figure 5, p. 968).

However, this boost to morphological rule learning is just a descriptive result that the original paper does not explore further. Additionally, this finding might come not from the production task *per se*, but rather from properties of the artificial language that Hopman and MacDonald used. The language was very complex in that every sentence contained multiple modifiers and adverbial phrases, so the word order rules might have been hard to identify. On the other hand, several words in every sentence contained identical suffixes, helping the morphological pattern stand out. Here, we aim to follow up on Hopman and MacDonald's result using an artificial language designed to tease apart adults' learning of morphological rules and word-level rules, and investigate how differential learning of these rules may depend on task.

Why would we expect language production to help adults learn morphological rules at all? First, it's useful to understand what makes learning morphological rules difficult in the first place. One of the major challenges, according to Ellis (2022), is their low perceptual salience. Morphemes tend to be smaller, shorter, unstressed, harder to segment, and less reliable in form than lexical units (DeKeyser 2005; Kenanidis et al. 2023; Sagarra & Ellis 2013). All of these things make them harder to notice, and thus harder to learn.

But one of the reasons that production has been suggested to strengthen language

learning is what Swain (2005) describes as its "noticing role". The idea is that production forces learners to pay more attention to the utterances they're assembling. And increased attention is associated with improved learning (Kenanidis et al. 2023; Schmidt 2001): this deeper processing makes learners more likely to notice linguistic patterns and induce possible generalisations (see also Izumi 2002). As long as the task is more active than a recognition-based comprehension task, we would expect the noticing role of production to take effect; based on literature on the effects of different kinds of tests, any kind of test beyond passive recognition should improve learning (see, e.g. Kang et al. 2007; McDaniel et al. 2007; McDermott et al. 2014). We therefore hypothesised that a production task would help people notice low-salience morphological patterns that they may otherwise have missed.

As a testing ground for this hypothesis, we used the well-studied trade-off between case marking, an example of a morphological rule, and fixed word order, an example of a word-level rule (Bentz & Winter 2013; Fedzechkina et al. 2011; Levshina 2020; Lupyan & Dale 2010). The rest of this chapter discusses two preregistered experiments (https://osf.io/qbjda) that test this hypothesis on two populations that differ in their prior experience with case-marking systems: Experiment 1 tests L1 English participants, and Experiment 2 tests L1 German participants.

To foreshadow our results: overall, participants learned the fixed word order rule but failed to acquire the case marking rule, although the majority did notice the recurring syllable pattern that was the consequence of case marking. Even participants already familiar with the concept of case (the German L1 participants in Experiment 2) showed the same clear preference to treat words as the smallest unit in the language and not to segment below this level. With respect to our main hypothesis, we found no evidence that taking part in a production task made participants in either experiment more likely to learn the case marking rule.

3.2 Experiment 1

Participants were trained on a series of sentences that each described a transitive event between two human characters. These sentences were designed to be compatible with

both word-level and morphological strategies for marking thematic role. Specifically, each sentence had the same fixed word order (SOV), a consistent word-level cue, *and* each noun bore a suffix corresponding to its grammatical role (nominative for the agent role and accusative for the patient role), a consistent morphological cue. In this way, the cues were not in direct competition: both were always available, and both reliably signalled the correct interpretation (E. Bates & MacWhinney 1981).

For example, participants might see an image of a fairy pushing a doctor and learn the corresponding sentence *fuvu zijo gix*. Then they might see a cowboy kicking a pirate and learn the sentence *lovu wujo kuv*. In both sentences, the word order is SOV, and in both sentences, the agent is marked with *-vu* and the patient with *-jo*. Thus participants could analyse the language in two different ways: like (1), in which nouns remain unsegmented, or like (2), in which nouns are segmented into stem and case marker.

(1) a. fuvu zijo gix
fairy doctor push

b. lovu wujo kuv
cowboy pirate kick

(2) a. fu-vu zi-jo gix
fairy-NOM doctor-ACC push

b. lo-vu wu-jo kuv
cowboy-NOM pirate-ACC kick

Note that the recurring syllables at the end of each noun could also be analysed in terms of their linear order: participants could arrive at an analysis like "the first noun always ends in *vu*, and the second noun always ends in *jo*". This is not a case marking analysis *per se*, since it's not based on thematic roles. But it is still of interest to us, because we're concerned with how well participants can identify patterns below word level. For this reason, in what follows, we refer to the two possible analyses not as "fixed word order" and "case marking", but rather as "unsegmented" and "segmented", respectively.

A crucial aspect of the training phase's design is that participants received no direct evidence that nouns have morphological structure, because none of the characters appeared as both agent and patient. Thus, the language's grammar is ambiguous. To

illustrate concretely: a participant would only ever see the fairy character as an agent, only ever labelled as *fuvu*. They receive no information about what form this word would take if the fairy were a patient. The word might become *fujo*, following the segmented analysis, or remain *fuvu*, following the unsegmented analysis. Thus it was possible for participants to successfully learn the training data without segmenting the words. We designed the language to be ambiguous because we wanted to know whether different kinds of practice task would help learners pick up on information that was consistently available, but not required to succeed at learning.

After training, we split participants into two groups to introduce the manipulation by task. Half of the participants practised the sentences they had learned using a more active production-like task (the PRODUCTION condition), while the other half practised using a more passive comprehension task (the COMPREHENSION condition). In the PRODUCTION condition, participants were required to actively construct sentences by clicking on the component syllables in the correct order, while the task in the COMPREHENSION condition simply involved choosing the correct image from an array of two.

Next, in the testing phase, we showed participants the same scenes they saw in training, but with the characters' roles reversed. For example, where in training they saw a fairy pushing a doctor, now they saw the doctor pushing the fairy. We then asked them to judge two different sentences that might describe this scene. The first kind of sentence was formed using the unsegmented analysis: the full words for the agent and patient were rearranged. The second kind of sentence was formed using the segmented analysis: only the stems were rearranged, and the case markers stayed in place.

If participants learned the nouns as unsegmented, holistic chunks, they should accept the first kind of sentence. If they segmented the nouns into stem and suffix, they should accept the second kind of sentence. Given previous findings that adults struggle to learn case morphology, we expected our participants to prefer sentences formed using the unsegmented analysis. However, crucially, here we test whether this preference is affected by the type of practice task they did.

Finally, participants completed a one-shot cloze task with a novel character i.e. a

character held out from the set encountered in training. The goal here was to assess whether participants were aware of the language's morphological patterns (that the first noun always ends in a particular syllable, and that the second noun always ends in another), whether or not they actually analysed these syllables as case markers.

3.2.1 Materials

The artificial language contained transitive sentences made up of three words: one for the agent, one for the patient, and one for the action, in that order (i.e. SOV). All verbs were monosyllabic CVC nonsense words, and all nouns were disyllabic CVCV nonsense words. Verbs were randomly selected from a set of 28: *gax*, *gix*, *gox*, *hix*, *jeg*, *jix*, *juf*, *juz*, *kex*, *kez*, *kuv*, *kux*, *nuz*, *puv*, *pux*, *vaf*, *vof*, *wez*, *wox*, *zax*, *zok*, *zox*, *zud*, *zuf*, *zug*, *zup*, *zuv*, and *zux*. Nouns were randomly assembled from nine possible stem syllables (*bu*, *fu*, *gu*, *ki*, *lo*, *ru*, *wu*, *ze*, and *zi*) and two suffix syllables (*vu* and *jo*) such that all agent nouns took one suffix and all patient nouns took the other.

Each sentence accompanied an image, a line drawing of two human characters interacting. A few examples are shown in Figure 3.1. The nine possible characters were: a chef, a cowboy, a doctor, a fairy, a footballer, a nun, a pirate, a princess, and a wizard. Each scene showed the agent character engaging in a reversible transitive action toward the patient character. The nine possible actions were: admiring, greeting, kicking, kissing, patting, poking, pushing, seeing, and yelling. Each image had two mirrored versions: one with the agent on the left, and one with the agent on the right.

To keep the artificial lexicon learnable, we randomly selected only six characters and two actions for each participant. The characters and actions were randomly associated with nonsense noun stems and verbs from the sets listed above. Then, each character was mapped to the thematic role they would appear in during the training phase. The mapping between characters and roles was random, with one constraint: we disallowed any permutations in which all agents were female and all patients were male (or vice versa), to forestall analyses of the suffixes as gender markers. All in all, participants saw 18 unique scenes during training: 3 agents × 3 patients × 2 verbs.

3.2.2 Procedure

We wrote the experiment in JavaScript using the jsPsych library (Leeuw et al. 2023). It contains four phases, detailed below and illustrated in Figure 3.1.

3.2.2.1 Training

In each training trial, participants saw an image alone for 1000 ms. Then the corresponding sentence in the artificial language appeared below it. 2500 ms later, a 'next' button appeared below the sentence. Clicking on it advanced participants to the next trial.

The whole training phase consisted of three blocks of 18 trials each, one trial per scene. Participants could optionally take a short break between each block.

3.2.2.2 Practice

After training, participants were divided into two groups: one group completed a more active production-like practice task (the PRODUCTION condition), and the other completed a more passive comprehension practice task (the COMPREHENSION condition). Both practice tasks involved familiar scenes and sentences that participants had encountered during training.

Participants in the PRODUCTION condition saw a familiar scene and had to build the correct sentence for this scene out of its component syllables. Below the image were five gaps, and below the gaps was one button per syllable in the sentence, shown in a random order. Although this task is less active than, say, speaking the artificial language sentence aloud, it still involves reproducing the linguistic signal that participants received. This reproduction places additional demands on participants that the comprehension task, as a simple recognition task, does not (more detail on the comprehension task below).

Clicking one of the buttons added that syllable into the leftmost gap in the sentence, so the sentence was filled in left to right as each syllable was clicked. An 'undo' button emptied the most recently filled gap. Participants could submit their sentence with the

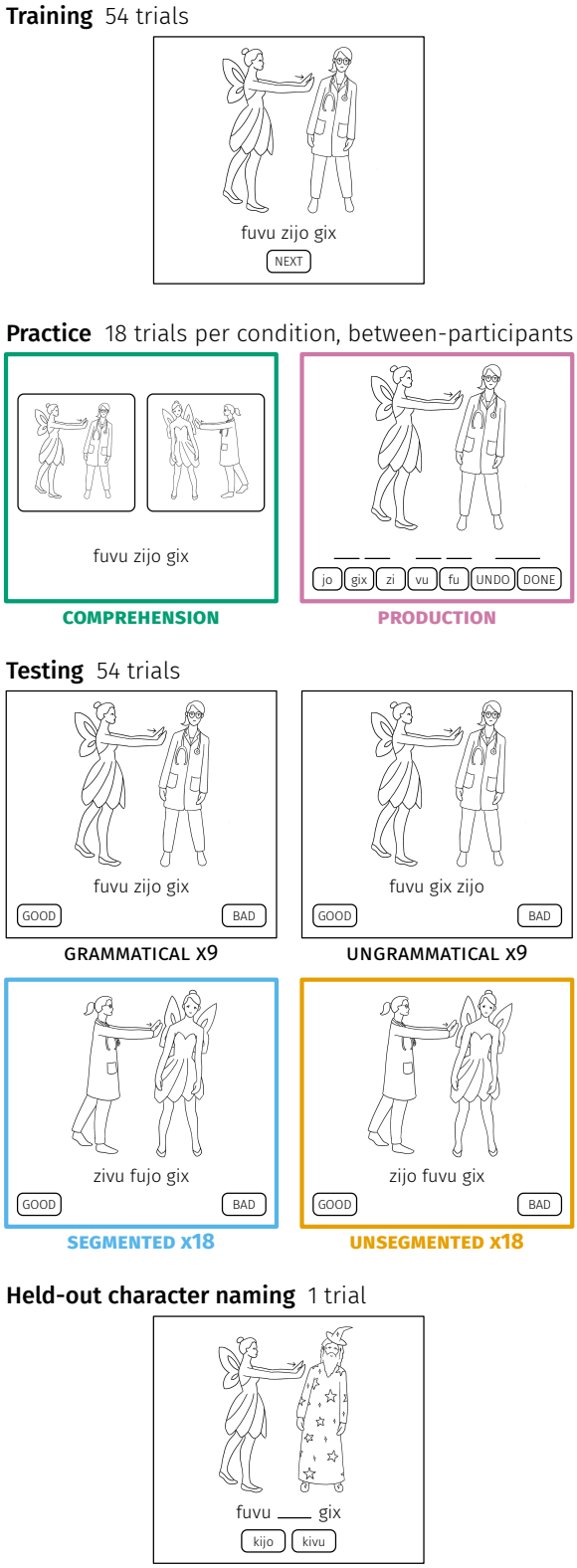

Figure 3.1: A schematic overview of trials in each of the experiment's four phases. All participants do the same training, then do either the comprehension or the production practice task. Then all participants complete the same testing and character naming phases. In other words, the two conditions differ only in the practice task. The colours of different trial types in this graphic match the colour coding used throughout this chapter.

'done' button as long as the sentence included every syllable once.

After submitting the sentence, participants received feedback on their response and were shown the correct sentence. The feedback stayed on-screen until participants clicked 'next' to continue. Each participant did 18 production trials, one per familiar scene, shown in a random order.

Participants in the COMPREHENSION condition were shown a familiar sentence and had to select the corresponding scene from an array of two. The target scene was a familiar one encountered during training; the foil image contained the same characters but with the thematic roles reversed (that is, if the target showed the fairy pushing the doctor, then the foil would show the doctor pushing the fairy). The order of target and foil was randomised on each trial. The agent appeared on the left in one image and on the right in the other, so that the characters themselves remained in the same position in each image.

So that we do not confound our results by giving participants in the PRODUCTION condition a segmentation advantage (in that they see each syllable individually on its own button), we made the sentence in the comprehension task appear on screen one syllable at a time, with a new syllable appearing every 500 ms. Once the full sentence was visible, participants could click on one of the two scenes. They received feedback on their response which stayed on-screen until they clicked 'next' to move to the next trial. Each participant did 18 comprehension trials, one per familiar scene, shown in a random order.

3.2.2.3 Testing

After the practice phase, all participants were asked to judge a number of sentences, some familiar and some novel. In each trial, participants saw a scene and a sentence, along with the prompt "Could someone who speaks this language describe this scene using the sentence below?". We used the `f` and `j` keys for "yes" and "no", with the mapping randomly determined for each participant (but kept the same for each trial). Participants received no feedback during this phase: pressing either `f` or `j` immediately moved them on to the next trial.

The testing phase contained four kinds of trials. First, there were GRAMMATICAL trials: nine of the familiar scenes and sentences from training, randomly sampled. If participants learned the language, they should always accept these sentences. Second, there were UNGRAMMATICAL trials: the other nine familiar scenes from training, but with sentences rearranged into a different word order (SVO, rather than the SOV participants were trained on). If participants learned the word order rule in the language, we reasoned that they should always reject these sentences.

We preregistered a particular exclusion criterion for these sentences which allowed participants to make up to and including four mistakes across these 18 GRAMMATICAL and UNGRAMMATICAL trials. In other words, the minimum accuracy permitted was 77.7%. However, it is worth noting that by excluding participants who accepted a different word order, we might be rejecting exactly those participants who had adopted a case marking analysis, since case marking languages generally permit freer word order (Fedzechkina et al. 2011; Lupyan & Dale 2010). In an exploratory analysis reported in Appendix 3.A, we removed the ungrammaticality criterion and re-ran the analysis we describe below, this time including participants who accepted any number of "ungrammatical" sentences. The overall pattern of results remains the same regardless of whether we use this criterion. This suggests that participants who accept the "ungrammatical" SVO sentences do so not because they have learned a free word order along with a case marking rule, but because they haven't learned the language reliably.

The final two trial types in the testing phase provide the critical data for our research question. In both trial types, the scenes contained familiar characters, but their thematic roles are reversed from the ones participants saw them in during training. Reversing the thematic roles causes the segmented analysis to yield a different sentence than the unsegmented analysis.

To illustrate: if a participant learned that the sentence *fuvu zijo gix* goes along with the fairy pushing the doctor, then in the testing phase, they would encounter two trials with a scene of the doctor pushing the fairy. In the SEGMENTED trial, they would see the doctor pushing the fairy along with the sentence in (3), which was formed by swapping just the CV stems. In the UNSEGMENTED trial, they would see this same scene along with the sentence in (4), which was formed by swapping the entire nouns.

Participants were asked to judge novel sentences formed according to these two rules for all 18 reversed-role scenes.

(3) zi-vu fu-jo gix
doctor-NOM fairy-ACC push

(4) zijo fuvu gix
doctor fairy push

All in all, the testing phase contained 54 trials (9 GRAMMATICAL + 9 UNGRAMMATICAL + 18 SEGMENTED + 18 UNSEGMENTED). The order of these trials was randomised for each participant.

3.2.2.4 Held-out character naming

The final phase of the experiment involved a one-shot trial in which participants saw a scene with one familiar character, one held-out character that had not been previously seen in the experiment, and a familiar action happening between them. These elements were all randomly chosen. The familiar character always appeared in the same thematic role from training, so the label for that character was also familiar. The held-out character assumed the other role.

Along with the scene, participants saw a sentence with a gap where the word for the new character would be. They were asked "What seems like the most plausible word for the new character in this scene?". Two alternatives were provided, formed by combining a random held-out stem with *-vu* and with *-jo*. For example, if the scene was the fairy (familiar noun) pushing the wizard (unfamiliar noun), and the sentence was *fuvu* ____ *gix*, participants might be asked to choose between *kivu* and *kijo* as the label for the wizard.

3.2.3 Participants and exclusions

We used Prolific to recruit 183 adults resident in the UK who self-reported that their first language was English and that they had no known language disorders. They all gave informed consent to participate in the experiment. The experiment was approved by the PPLS Ethics Committee at the University of Edinburgh (ref. 230-2223/2).

The experiment took around 20 minutes to complete (median time = 17:38), and participants were paid £3.50 (above UK National Minimum Wage at the time of running the experiment). Participants were randomly assigned to either the COMPREHENSION condition (100 people) or the PRODUCTION condition (83 people). We excluded 103 participants for the following preregistered reasons: self-reporting the use of written notes in an exit questionnaire contrary to instructions (2); low accuracy (< 77.7%, i.e. 14/18) on practice trials (16), testing trials (49) or both (36).

Figure 3.2 illustrates how many times each exclusion criterion was met in each condition. This plot does not reflect how the criteria may overlap, so participants caught by multiple criteria contribute to multiple counts; see Appendix 3.B for a full breakdown of how many participants were caught by each combination of criteria.

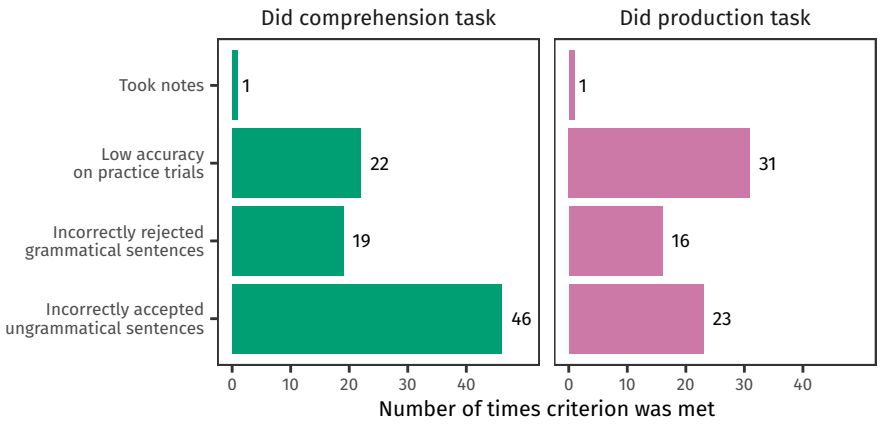

Figure 3.2: How many times each preregistered exclusion criterion was met in Experiment 1 (participants caught by more than one criterion contribute to each criterion's count). On the whole, exclusion criteria were met more often in the COMPREHENSION condition than in the PRODUCTION condition.

We had to exclude many more participants who had been originally recruited into the COMPREHENSION condition, and fewer who were recruited into the PRODUCTION condition. This asymmetry indicates at least anecdotally that, despite being a somewhat unnatural simulation of natural language production, our "production" task does seem to have helped participants learn the sentences that they were exposed to — in line with previous evidence that more active tasks are good for learning.

After exclusions, we were left with analysable data from 40 participants in each condition. (Appendix 3.C contains the same analysis that we report below run on all 183 participants.) Accuracy on the grammatical and ungrammatical sentences was close to ceiling for all remaining participants (naturally, since these are the participants

who were not excluded for low accuracy), and there were no substantial differences between conditions. For the COMPREHENSION group, grammatical sentences were correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 98% of the time. And for the PRODUCTION group, grammatical sentences were also correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 97% of the time.

3.2.4 Results

3.2.4.1 Judgement

Participants in both conditions tended to accept novel sentences formed using the unsegmented analysis, and they were more ambivalent about novel sentences formed using the segmented analysis. Figure 3.3 shows the proportion of each kind of novel sentence that participants accepted.

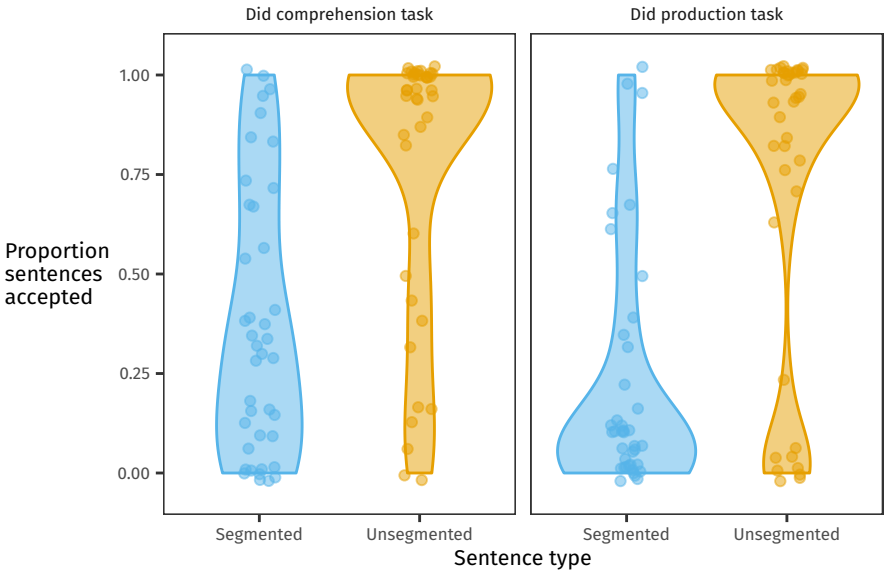

Figure 3.3: Participants in both the COMPREHENSION and PRODUCTION conditions of Experiment 1 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis. Each dot represents one participant's proportion of accepted sentences of each type.

Following our preregistered analysis plan, we used `brms` (Bürkner 2017) in R (R Core Team 2024) to fit a Bayesian linear model with a Bernoulli likelihood to this data. This model predicts sentence acceptance as a function of condition (COMPREHENSION

vs. PRODUCTION), sentence type (SEGMENTED vs. UNSEGMENTED), and their interaction. The group-level effects in the model included varying intercepts by participant and varying slopes over sentence type by participant. We selected the model's weakly regularising priors using prior predictive checks (for more detail, see Appendix A in Pankratz 2025). The model converged, as indicated by all Rhats = 1.00. Appendix 3.D contains the full model specification.

We sum-coded condition (COMPREHENSION as –0.5, PRODUCTION as +0.5) and sentence type (SEGMENTED as –0.5, UNSEGMENTED as +0.5). The interaction term was also scaled to ±0.5 so that we could use the same weakly regularising prior for all three predictors.

We hypothesised that, if a production task helps participants learn morphological rules, then participants in the PRODUCTION condition would be more likely to accept sentences generated by the segmented analysis than participants in the COMPREHENSION condition. We would see this in the model as a high-certainty interaction between condition and sentence type.

The model's posterior estimates for the population-level effects are summarised in Table 3.1. Figure 3.4 shows the conditional posterior probability distributions — that is, the posterior distributions over the probabilities of accepting a sentence for all combinations of condition and sentence type.

Table 3.1: The posterior probability distributions estimated by the model for the English participants' sentence acceptance data in Experiment 1. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.54	0.27	0.03	1.10
Condition	–0.81	0.51	–1.79	0.20
Sentence type	4.10	0.70	2.76	5.51
Condition:Sent. type	0.37	0.66	–0.88	1.70

Overall, the model indicates with high certainty that participants are more likely to accept a novel sentence formed with the unsegmented analysis compared to a novel sentence formed with the segmented analysis. Concerning condition, the model's estimates indicate uncertainty about a difference in sentence acceptance probabilities between the PRODUCTION condition and the COMPREHENSION condition, as well as uncertainty about the interaction that our hypothesis targeted. Our prediction that participants who did the production task would be more likely to accept SEGMENTED

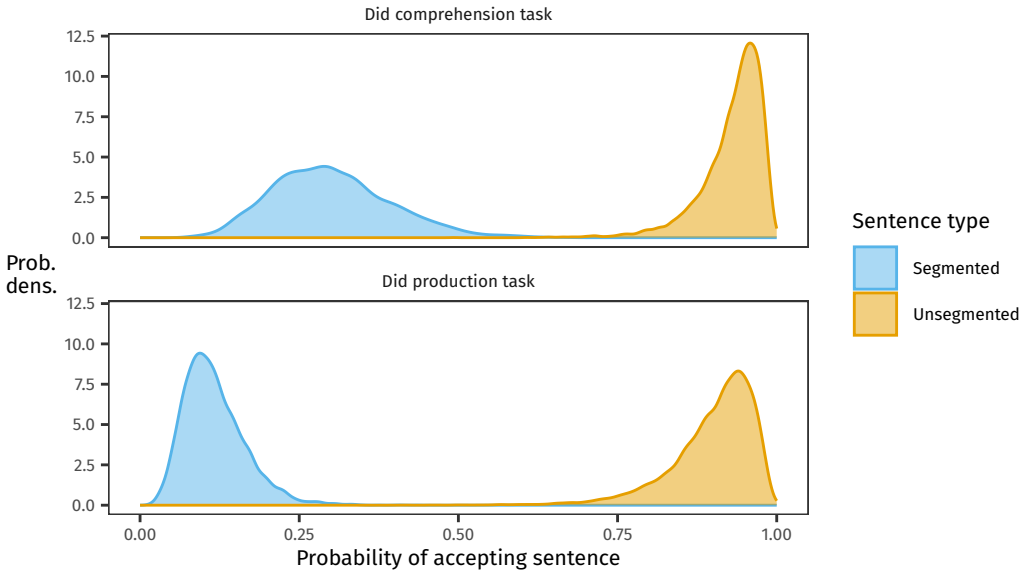

Figure 3.4: Conditional posterior probability distributions of the probability of accepting a sentence in Experiment 1. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

sentences was not borne out, and in fact, the results trend slightly in the opposite direction.

3.2.4.2 Held-out character naming

Figure 3.5 shows the proportion of participants who chose a label for the held-out character that contained the appropriate suffix. Even if they didn't arrive at a fully-fledged case marking analysis, more than half of the participants in each condition seem to have noticed that each noun reliably ends in a particular syllable.

We preregistered the analysis of this data as exploratory. To see whether participants in the PRODUCTION condition showed greater awareness of these morphological patterns (even if they did not analyse them as case markers *per se*), we fit a Bayesian linear model with a Bernoulli likelihood to this data, predicting appropriate suffix choice as a function of condition (COMPREHENSION coded as –0.5, PRODUCTION as +0.5). Every participant gave only one data point, so no group-level effects were needed.[1] We used the same weakly regularising priors as in other Bernoulli models reported in this

[1]Because we only analyse one data point for each participant, we implicitly assume that there is no variability in the responses to this task. Naturally, the absence of variability is very unlikely, and this assumption is a limitation of the one-shot trial format. This format was simply an inexpensive way to gather exploratory data, which we hoped might open the door to studying a potentially interesting question in greater depth in the future.

chapter. The model converged, as indicated by all Rhats = 1.00.

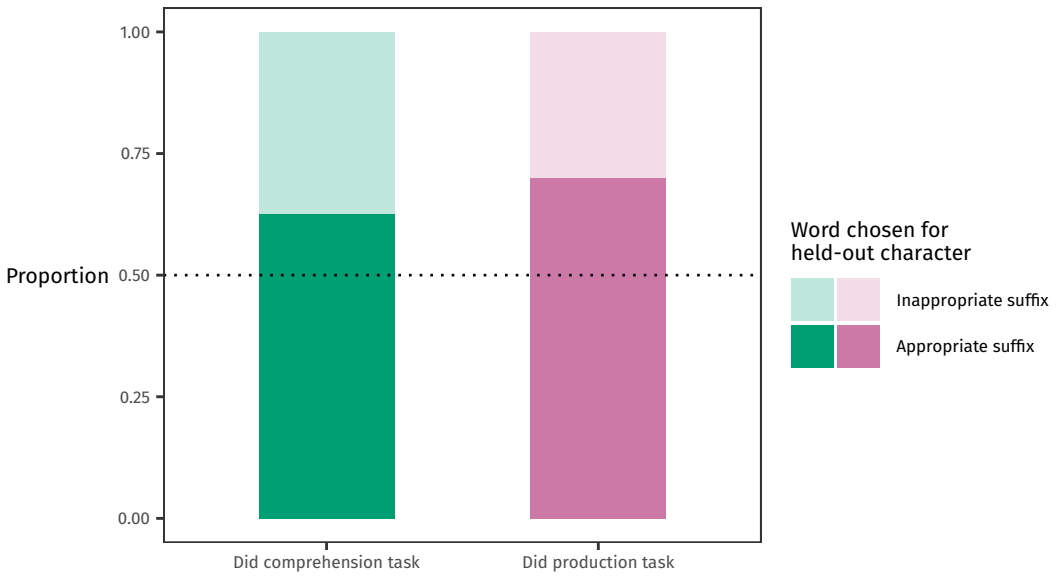

Figure 3.5: In the held-out character naming task of Experiment 1, more than half of participants in each condition selected the word with the appropriate suffix. Slightly more participants in the PRODUCTION condition selected the appropriate suffix.

Table 3.2 summarises the posterior distributions of the population-level effects estimated by this model, and Figure 3.6 shows the conditional posterior probability distributions over the probabilities of selecting the appropriate suffix.

Table 3.2: The posterior probability distributions estimated by the model for the English participants' held-out character naming data in Experiment 1. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.68	0.24	0.23	1.16
Condition	0.33	0.47	–0.57	1.22

The model indicates that participants in both groups chose the label containing the appropriate suffix for the missing word with a probability slightly greater than chance. Although being in the PRODUCTION condition is associated with a slightly higher probability of choosing the appropriate label, there is a great deal of overlap between conditions and thus a great deal of uncertainty about whether participants in either condition are more likely to select the appropriate label.

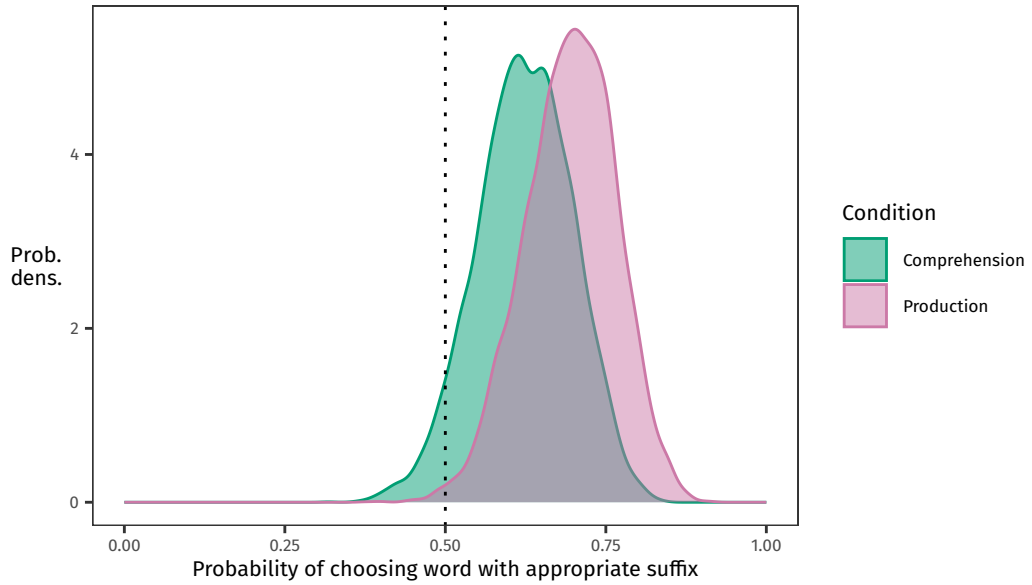

Figure 3.6: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 1. The overlap of these posteriors suggests uncertainty about whether and how much the groups might differ.

3.2.5 Interim discussion

Experiment 1 has shown that participants in both groups overwhelmingly preferred novel sentences formed using the unsegmented analysis over sentences formed using the segmented analysis. This preference was not clearly affected by the type of task participants used to practise the language, counter to our hypothesis.

Interestingly, the preference for the unsegmented analysis was resounding, even though the held-out character naming task indicated that many participants were aware of a morphological pattern in the language — namely that the two nouns always ended in different syllables. That said, we cannot identify from this task whether learners were aware that these syllables also had a consistent linear order in the sentence, which would be a prerequisite for arriving at the segmented analysis.

One straightforward explanation for why L1 English participants might identify the morphological pattern without learning the segmented analysis is that case is not morphologically marked outside of the pronominal system in English. In other words, English uses word order alone to indicate grammatical roles, and thus our participants may have been particularly unlikely to look beyond word order to notice that the case marking suffixes also indicated these roles (see similar observations in in Kenanidis et al. 2023). We collected data about further languages that participants know or under-

stand, and, in an exploratory analysis, compared the performance of participants who do know a case-marking language (15 people) to those who do not (65). The pattern of results remains the same; see Appendix 3.E for details.

Nonetheless, it is possible that a population whose L1 includes more widespread use of case would be more likely to access the case marking analysis. We therefore ran a follow-up experiment with L1 speakers of German, a language with a productive case marking system featuring (among other cases) nominative and accusative differentially marked on nominal dependents like determiners.

German speakers also have plenty of experience to suggest that word order is not always a reliable cue to thematic role, since German is a V2 language with no restrictions on the grammatical category of the preverbal constituent (Holmberg 2015; Meisezahl et al. 2023). They might therefore be more willing than English participants to look beyond word order and notice additional generalisations available in the language. In line with these predictions, Kenanidis et al. (2023) found that although L1 German participants learned case marking rules less well than word order rules, they were better at learning the case marking rules than L1 English participants were.

3.3 Experiment 2

3.3.1 Materials

We used largely the same materials as in Experiment 1, described above in Section 3.2.1. Only a handful of changes were made for German-speaking participants.

First, we removed any forms from the language that resembled German words: *zug* is like German *Zug* 'train', *kex* might be read as *Keks* 'cookie', and so on.

Second, to ensure that the full set of stimuli was grammatically equivalent in German, we removed all images containing the pirate character. The German word *Pirat* is a so-called "strong masculine" noun: a noun that itself inflects for case, in addition to the usual inflection on the determiner (cf. nominative *der Pirat* 'the pirate', accusative *den Piraten*). All other characters correspond to German nouns that are grammatically "weak", that is, the nouns don't inflect for case.

Third, we changed the default word order from SOV to VSO, because SOV is sometimes argued to be the basic word order of German (Haftka 1996; Haider 2020). This means that the Experiment 1 sentence *fuvu zijo gix* would become *gix fuvu zijo* in Experiment 2. The "ungrammatical" word order in the judgement phase remained SVO, akin to German's V2 (though we did not use rejection of SVO sentences as a criterion for excluding participants in Experiment 2; we will discuss this further in Section 3.3.3).

3.3.2 Procedure

Experiment 2 followed the same procedure as Experiment 1 (see Section 3.2.2), with one modification. For English participants, we had randomly mapped the keys `f` and `j` to 'yes' and 'no'. Since German *ja* 'yes' begins with J, we instead used `p` and `q` as the decision keys for the sentence judgement task.

3.3.3 Participants and exclusions

We used Prolific to recruit 135 participants who self-reported that their first language was German and that they had no known language disorders. They all gave informed consent to participate in the experiment. The experiment was approved by the PPLS Ethics Committee at the University of Edinburgh (ref. 230-2223/4).

The experiment took around 20 minutes to complete (median time = 17:39), and participants were paid £3.85 (approx. €4.50), above UK National Minimum Wage at the time of running the experiment. As in Experiment 1, participants were randomly assigned to either the COMPREHENSION condition (68 people) or the PRODUCTION condition (67 people). We excluded 43 participants for the following preregistered reasons: low accuracy on practice trials (17), GRAMMATICAL testing trials (12), or both (14). Figure 3.7 illustrates how many times each exclusion criterion was met in each condition (note that this plot does not reflect how criteria may overlap, so participants caught by multiple criteria contribute to multiple counts).

This figure includes German participants' performance on the so-called "ungrammatical" sentences, the ones with word order that differs from training, though we did not use this criterion to exclude participants from the analysis. We ignored this cri-

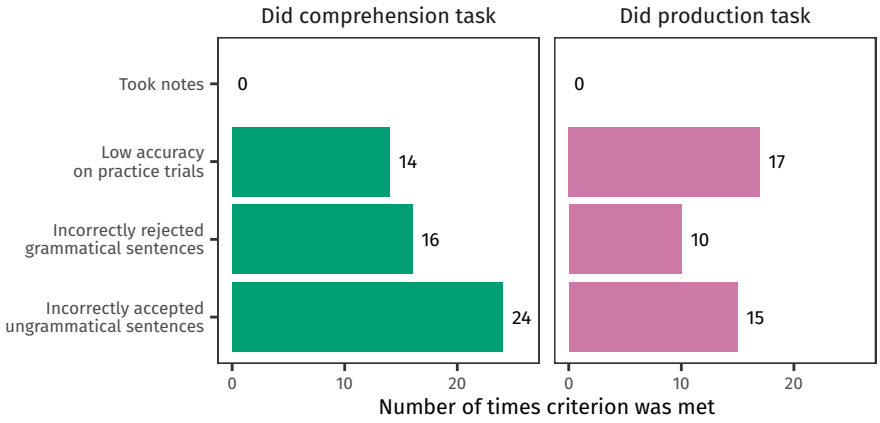

Figure 3.7: How many times each preregistered exclusion criterion was met in Experiment 2 (participants caught by more than one criterion contribute to each criterion's count). Exclusions are more balanced between conditions in Experiment 2 compared to Experiment 1, though still, more participants in the COMPREHENSION group compared to the PRODUCTION group incorrectly rejected familiar grammatical sentences. (The ungrammatical sentences criterion is included here only for completeness; in Experiment 2 it was not used to exclude participants.)

terion for German speakers because German permits a relatively free word order, so participants may not have had the expectation that word order should be fixed, particularly if they accessed the segmented (case marking) analysis. Recall that removing this criterion for the English-speaking participants in Experiment 1 did not affect the pattern of results (Appendix 3.A).

After exclusions, we were left with data from 46 participants in each condition. The remaining participants' accuracy on the grammatical and ungrammatical sentences was fairly high, with no substantial differences between conditions. For the COMPREHENSION group, grammatical sentences were correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 78% of the time. And for the PRODUCTION group, grammatical sentences were also correctly accepted 96% of the time, and ungrammatical sentences were correctly rejected 82% of the time.

3.3.4 Results

Overall, the results from the German participants in Experiment 2 are similar to the results from the English participants in Experiment 1.

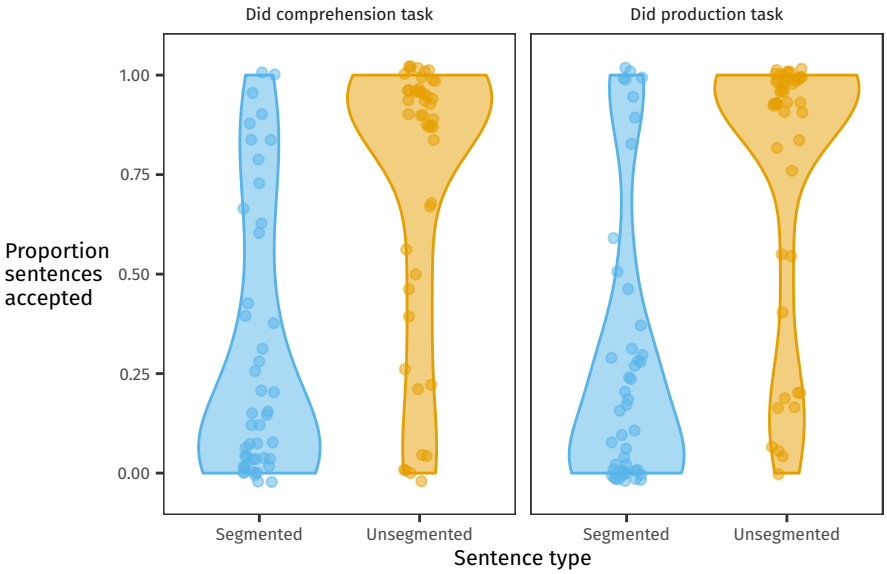

Figure 3.8: In Experiment 2, participants in both the COMPREHENSION and PRODUCTION conditions again accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis.

3.3.4.1 Judgement

Like the English-speaking participants, the German participants showed a clear preference for the unsegmented analysis (see Figure 3.8). We fit the same Bayesian linear model as described above in Section 3.2.2.3 to the data from the German participants. The posterior distributions for the population-level effects estimated by the model are given in Table 3.3, and the conditional posterior probability distributions are shown in Figure 3.9. Again, we cannot be certain about the interaction that would support our hypothesis about a production task enabling participants to learn the segmented analysis.

Table 3.3: The posterior probability distributions estimated by the model for the German participants' sentence acceptance data in Experiment 2. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.17	0.21	–0.24	0.59
Condition	0.33	0.40	–0.45	1.13
Sentence type	3.84	0.59	2.68	5.01
Condition:Sent. type	0.52	0.58	–0.61	1.68

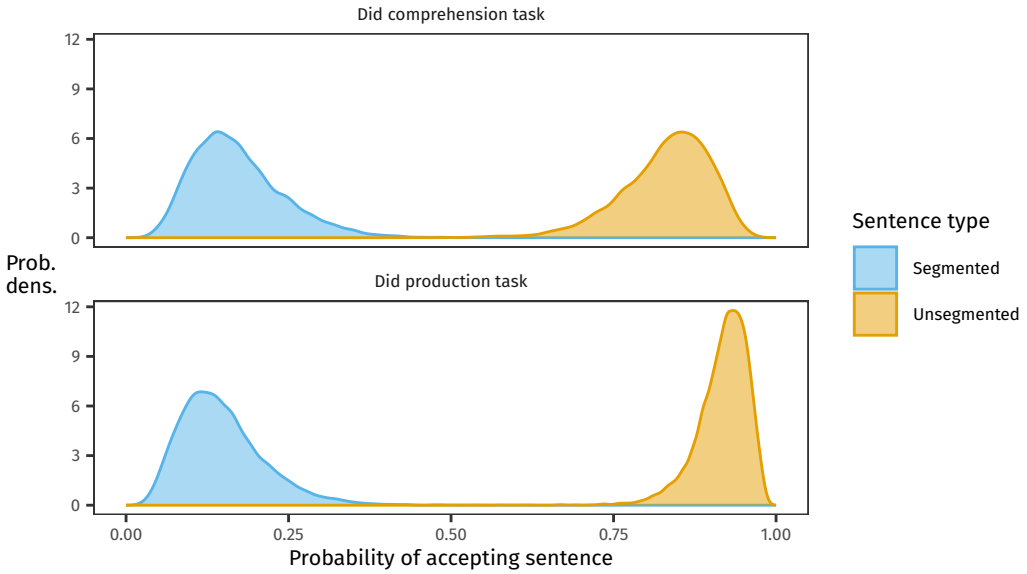

Figure 3.9: Conditional posterior probability distributions of the probability of accepting a sentence for the participants in Experiment 2. As in Experiment 1, UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

3.3.4.2 Held-out character naming

About three-quarters of German participants appear to have noticed that one noun always ends in *-vu* and the other always ends in *-jo*; see Figure 3.10.

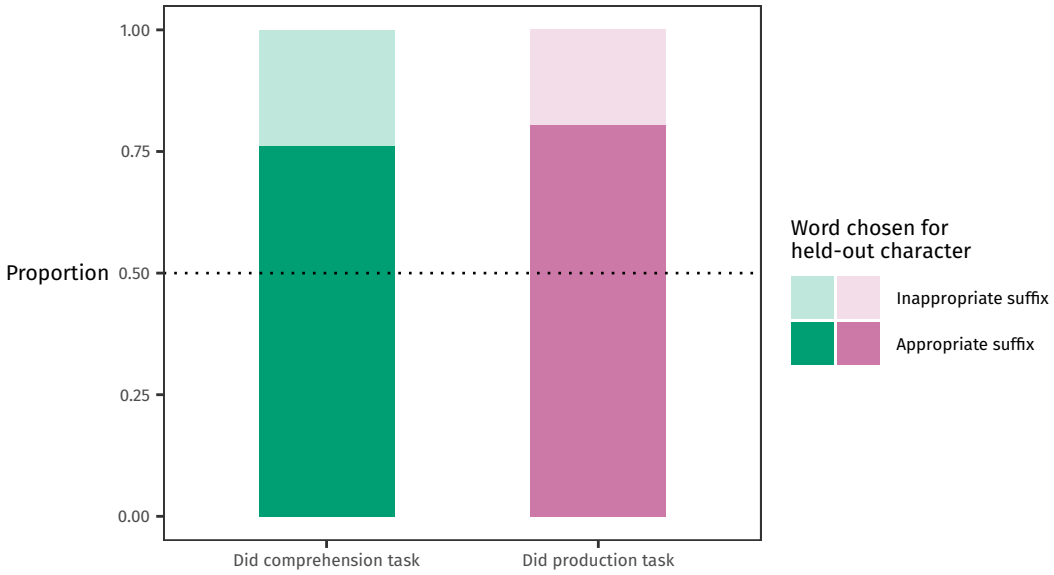

Figure 3.10: In the held-out character naming task of Experiment 2, around three-quarters of German participants selected the form in which the word ended in the appropriate suffix; the proportion of appropriate choices is slightly higher in the PRODUCTION condition.

We fit the same model as described in Section 3.2.4.2 to this data. Table 3.4 sum-

marises the posterior distributions of the population-level effects, and Figure 3.11 shows the conditional posterior probability distributions over the probabilities of selecting the appropriate suffix. The model suggests that, much like the English participants, the German group is likely to have labelled the held-out character following the morphological pattern, and there is no clear association between participants' choice of label for the held-out character and experimental condition.

Table 3.4: The posterior probability distributions estimated by the model for the German participants' held-out character naming data in Experiment 2. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	1.28	0.25	0.80	1.79
Condition	0.24	0.50	–0.73	1.22

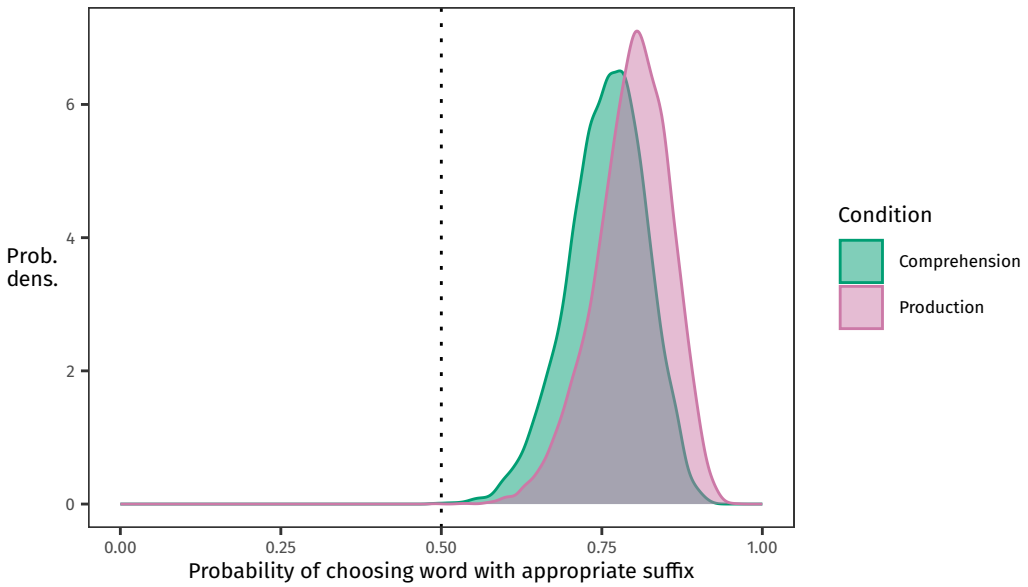

Figure 3.11: Conditional posterior probability distributions over the probability of German participants selecting a word that contains the appropriate suffix in Experiment 2. These posteriors overlap, so we are not certain whether and how much the groups might differ.

3.4 Combined analysis of Experiments 1 and 2

Finally, to see whether there are any meaningful differences between the English L1 speakers in Experiment 1 and the German L1 speakers in Experiment 2, we pooled the data from the two experiments and reran the main statistical models described above with one additional predictor: language. We sum-coded language on the same scale as all other predictors (English as -0.5, German as +0.5) to enable us to use the same

weakly regularising prior for all predictors.

3.4.1 Judgement

We fit a Bayesian linear model with a Bernoulli likelihood to the combined data. This model predicts sentence acceptance as a function of condition (COMPREHENSION vs. PRODUCTION), sentence type (SEGMENTED vs. UNSEGMENTED), language (English vs. German), and all two-way and three-way interactions. The model converged, as indicated by all Rhats = 1.00. The model's posterior estimates for the population-level effects are summarised in Table 3.5. Figure 3.12 shows the conditional posterior probability distributions — that is, the posterior distributions over the probabilities of accepting a sentence for all combinations of condition, sentence type and language.

Table 3.5: The posterior probability distributions estimated by the model for all participants' sentence acceptance data across the two experiments. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.38	0.16	0.06	0.70
Condition	-0.21	0.31	-0.82	0.41
Sentence type	4.14	0.45	3.29	5.03
Language	-0.21	0.31	-0.82	0.39
Condition:Sent. type	0.95	0.83	-0.65	2.56
Sent. type:Language	-0.17	0.84	-1.78	1.46
Condition:Language	1.15	0.59	0.00	2.30
Condition:Sent. type:Language	0.15	1.32	-2.48	2.73

Overall, the model indicates with high certainty that participants across the board are more likely to accept a novel sentence formed with the unsegmented analysis compared to a novel sentence formed with the segmented analysis. However, the model also reveals an interaction between condition and language, whereby the effect of practice condition is slightly stronger for English participants than German participants. Qualitatively, this interaction is capturing the observation that English participants showed a marginally stronger preference for the unsegmented analysis in the PRODUCTION condition than in the COMPREHENSION condition (as indicated by a wider posterior distribution over the probability of accepting a SEGMENTED sentence in the COMPREHENSION condition: Figure 3.12, blue distribution in the top-left panel compared to the same distribution in the top-right panel), while for the German participants, the strength of this preference was similar across practice conditions. In other

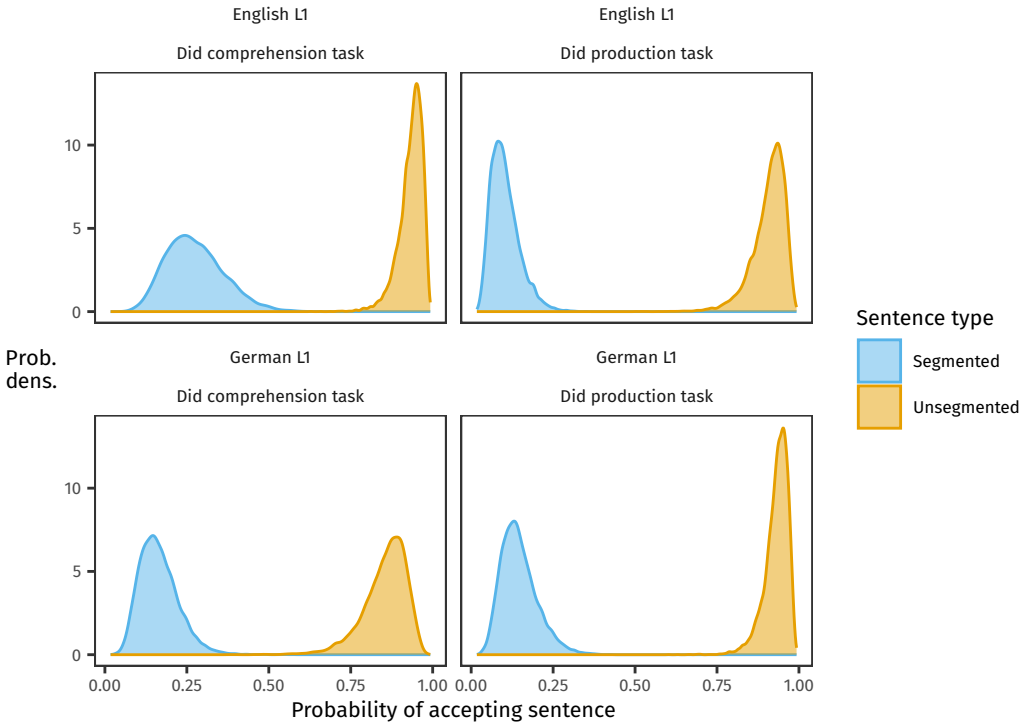

Figure 3.12: Conditional posterior probability distributions of the probability of accepting a sentence for all participants across the two experiments. In both experiments, UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task. However, English participants in the COMPREHENSION condition do show a slightly weaker preference for the UNSEGMENTED sentences than those in the PRODUCTION condition; this effect is less pronounced for German participants, whose responses are similarly polarised across practice conditions.

words, German participants showed more certainty about their chosen analysis, regardless of the type of practice task they completed.

3.4.2 Held-out character naming

We fit a Bayesian linear model with a Bernoulli likelihood to the combined data, predicting appropriate suffix choice as a function of condition (COMPREHENSION vs. PRODUCTION), language (English vs. German), and their interaction. The model converged, as indicated by all Rhats = 1.00.

Table 3.6 summarises the posterior distributions of the population-level effects estimated by this model, and Figure 3.13 shows the conditional posterior probability distributions over the probabilities of selecting the appropriate suffix.

The model indicates that, overall, participants chose the label containing the appro-

Table 3.6: The posterior probability distributions estimated by the model for all participants' held-out character naming data across the two experiments. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.99	0.18	0.65	1.34
Condition	0.29	0.35	-0.39	0.99
Language	0.61	0.35	-0.07	1.30
Condition:Language	-0.08	0.68	-1.39	1.25

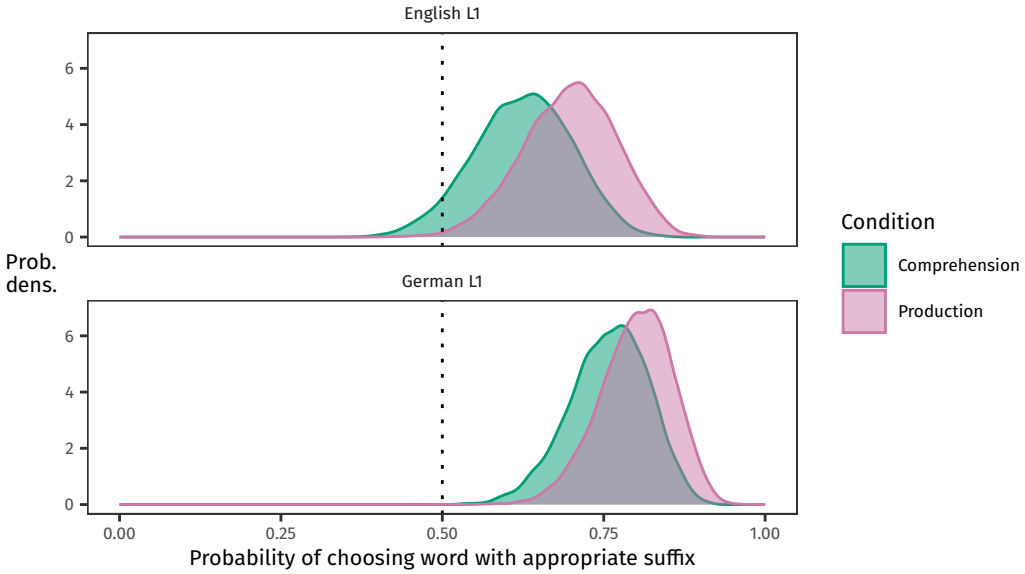

Figure 3.13: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix across the two experiments. The overlap of the posteriors suggests uncertainty about whether and how much the conditions might differ, although it does appear that German participants were overall more likely to choose the word containing the appropriate suffix.

priate suffix for the missing word with a probability greater than chance. Although being in the PRODUCTION condition is associated with a higher probability of choosing the appropriate label, there is considerable uncertainty about this effect. However, the model does indicate with moderate certainty that the German participants were overall slightly more likely than the English participants to select the label containing the appropriate suffix.

3.4.3 Discussion

Overall, comparison between the two experiments reveals some small but robust differences between the English and German participants in Experiments 1 and 2. Concretely, German participants in the COMPREHENSION condition gave more polarised responses to the two sentence types in the judgement task, reflecting greater certainty

that the unsegmented analysis was the correct one than the English participants in the same condition. And German participants were more likely to select the word containing the appropriate suffix in the held-out character naming task.

In a sense, both of these results might suggest that the German participants were slightly better learners than the English participants. In terms of the judgement data, this may seem like a counterintuitive suggestion at first glance, given that the German participants were no more likely than the English participants to infer a case marking rule from the artificial language — counter to our hypothesis. However, it could be argued that a lower willingness to segment the nouns is reflective of stronger lexical learning. In other words, the English participants in the COMPREHENSION condition may have been more ambivalent about the sentences following the segmented analysis because they simply hadn't learned the words as well, and were therefore less aware of the suffix changes. German participants' stronger performance in the held-out character naming task provides more unambiguous evidence that they internalised the language's regularities better than the English participants, albeit the effect size is small.

Although we can only speculate as to the origin of these effects, there are a variety of differences between the two populations that could plausibly affect performance on a language learning task. Firstly, the vast majority of the German participants (90 out of 92) were multilingual, compared to only 31 out of 80 English participants (eight of whom indicated that they had only a very basic level of proficiency in their additional languages). Thus, the German participants were clearly bringing more general language learning experience to the table. Secondly, German is a more morphologically rich language than English, so it is perhaps unsurprising that the German participants showed greater sensitivity to sub-word regularities in the held-out character naming task. Finally, and most speculatively, there may be differences in motivation between the two populations due to their experience on the Prolific platform. Anecdotally, where studies hosted on Prolific include requirements for language background, they are more often than not targeted at native English speakers. Therefore, it seems reasonable to assume that the German participants receive fewer opportunities to take part in studies. And since Prolific participants rely on meeting researchers' inclusion criteria to get paid, this scarcity of opportunity may encourage those in more niche

demographics to try harder to perform well. For native English speakers, on the other hand, the pressure to perform may feel lower, since they will always have access to a plethora of other studies if one doesn't go well. Moreover, if English-speaking participants are indeed taking part in more studies, they are more likely to experience respondent fatigue, which can lead to a decrease in data quality (e.g. Savage & Waldman 2008).

3.5 General discussion

3.5.1 Summary of results

In two artificial language learning experiments, we tested whether a production-like task — known to improve rule learning in a number of contexts — could also draw adult learners' attention to rules that adults typically overlook (Ellis 2022; Swain 2005). Specifically, we focused on morphological marking of thematic roles using case suffixes.

We trained participants on a language with fixed word order in which agent nouns and patient nouns were always marked with distinct suffixes (e.g. *-vu* for agents and *-jo* for patients). However, the nouns that each participant saw only ever occurred as either agents or patients, never in both roles. Thus the suffixes could be analysed as part of the nouns themselves (an unsegmented analysis) or as productive endings, part of a wider case system (a segmented analysis).

We found that regardless of whether participants did a production or a comprehension practice task, they favoured novel sentences which were formed using an unsegmented, word-level analysis, and they tended to reject sentences formed using a segmented, case-marking analysis. In other words, when shown novel scenes in which familiar characters featured in a novel grammatical role (e.g. where the fairy, which appeared only as an agent in training, appeared as the patient), they tended to reject sentences in which the noun suffixes were adjusted to reflect these new grammatical roles. Nevertheless, in the held-out character naming task, most participants showed that they detected the morphological pattern that resulted from the case marking (i.e.

that one noun in every sentence ended in *vu* and the other in *jo*), even if they did not necessarily develop this observation into a productive case marking grammar. Perhaps surprisingly, we found that the same pattern of results — sensitivity to the morphological patterns but failure to accept sentences formed according to the segmented analysis, regardless of practice condition — also held for participants whose first language, German, has extensive case marking. Reassuringly, though, similar results emerge in Kenanidis et al. (2023): both L1 English and L1 German participants were better at detecting word order violations compared to case marking violations.

In a sense, reanalysing an unsegmented word-order-based grammar into a case marking grammar is not a trivial task, since it means overriding the chunks that have already been learned. But it is something that learners of genuine case marking languages are likely to need to do — many nouns are more likely to occur in a particular grammatical role, e.g. humans and other animate beings are more commonly found as agents than as patients (Croft 2003; Meir et al. 2017; Silverstein 1976). So it is not unreasonable to expect our participants to be able to break down the chunks they have learned.

And indeed, some learners did seem to show forays in this direction. Our adult participants do "start big" (Christiansen & Chater 2016b; Havron & Arnon 2021; Siegelman & Arnon 2015; Wray 2002, 2006) by learning a rule that manipulates the larger, unsegmented units, not the smaller ones that require segmentation. But participants' behaviour when naming the held-out character adds some nuance to this result. Participants still notice patterns and pieces within the word-level chunks they manipulate, although contrary to our hypotheses, we observed no clear effect of task on this process. It appears that noticing is not enough — or perhaps the noticing was just too little, too late, as we'll discuss in Section 3.5.2 below.

A potential alternative explanation for participants' unwillingness to segment their learned chunks is that they were treating the suffixes not as case markers, but as markers of grammatical gender. In this case, participants would not expect a character's suffix to vary according to their semantic role in the sentence. Although our data cannot rule out this possibility, there are several reasons to be skeptical. Firstly, for our English-speaking participants, we would expect grammatical gender to be just as in-

accessible an analysis as case, since it is not a feature of their native language. Furthermore, we know that people's ability to learn noun classes depends heavily on the presence or salience of conditioning cues — neither adults nor children appear to readily acquire arbitrary subclass distinctions (Braine et al. 1990b; Culbertson & Wilson 2013; Frigo & McDonald 1998; K. H. Smith 1969) — and we took care to ensure that natural gender was not available as a cue to suffix assignment in the training data (see Section 3.2.1). Finally, if there was a group-level tendency to analyse the suffixes as gender markers, we would have expected performance on the held-out character naming to be closer to chance, since there would be no *a priori* reason to expect the two characters in a scene to be from different gender classes, and therefore no reason to assign them different suffixes. However, it is true that some of the participants who did not select the correct suffix during this task may have done so precisely because they had developed such an analysis — perhaps based on some other feature of the images that we did not control for e.g. whether or not the character had a hat.

Overall, though, we have not found a clear association between production practice and adults' ability to acquire a more difficult morphological rule when a more accessible word-level rule is also present. However, the present study has a number of issues that we believe stand in the way of being able to learn very much from our results. We review these next, and then turn in Section 3.5.3 to some suggested follow-up experiments that offer possible ways forward.

3.5.2 Limitations of the present study

We have identified three major limitations of the present study that must be remedied in future work: the task manipulation came too late in the learning process, the test sentences don't conclusively show us which cues participants learned, and the task we set for participants was just too difficult. We'll go through each of these issues one-by-one.

3.5.2.1 Placement of the task manipulation

In our view, the main reason that we failed to find an improvement with production is that participants in the PRODUCTION condition were not required to produce the language early enough in the learning process: the critical practice phase came only after an initial training phase. During this training phase, participants likely already discovered the fixed word order rule, and since that rule perfectly explained all the data they encountered, there was no need to search for further explanations (in classical conditioning terms, an *overshadowing* effect; Pavlov 1927).

This pattern of behaviour is characteristic of how adults approach many kinds of tasks, both linguistic and non-linguistic: they tend to identify a reliable cue and then exploit it. Children, in contrast, will notice a reliable cue but still continue to explore (Liquin & Gopnik 2022; Sumner et al. 2019). A possible prediction of this account, then, is that children might be more likely than adults to accept the case marking analysis.

The late start of the production/comprehension tasks could also be part of why our results differ from those of Hopman and MacDonald (2018), who observe that a production task leads to slightly better learning of morphological rules than word order rules. In their design, passive exposure trials were interleaved with blocks of active production trials. Interleaving training and testing like this is likely to improve performance by giving participants more chances to test their learning (Ambrose et al. 2010).

3.5.2.2 Design of the test stimuli

Secondly, the sentences we asked participants to judge in the test phase were not adequately designed to identify what analysis participants actually learned. We should have tested all combinations of correct and incorrect morphology and word order; the UNSEGMENTED trial type has correct word order but incorrect morphology, and the UNGRAMMATICAL trial type is only wrong with respect to the verb's location, not the relative order of agent and patient.

The main problem here is that none of the sentences we showed participants conflict with the word order cue of the agent appearing first and the patient second. In

other words, all of the sentences that participants saw throughout the experiment were consistent with the word order cue, but not all of them were consistent with the morphological cue. According to the Competition Model (E. Bates & MacWhinney 1981; MacWhinney 2018), this brings the two cues into conflict: specifically, word order becomes a more reliable cue to thematic role assignment than case marking. Assuming that participants continue to learn during the test phase, this competition between cues would strengthen learners' preference for the word order analysis.

Table 3.7 shows an example set of test sentences that would have better allowed us to tease apart participants' preferred cue(s).

Table 3.7: An improved set of test sentences derived from the example training sentence *fuvu zijo gix* 'fairy pushes doctor'. (Segmented components: *fu* 'fairy', *zi* 'doctor', *-vu* 'NOM', *-jo* 'ACC', *gix* 'push'). All of these sentences would describe the scene 'doctor pushes fairy'. Using sentences like these which cover all combinations of correct and incorrect morphology and word order would have allowed us to more accurately see which cues learners preferred.

	Correct word order			Incorrect word order		
Correct morphology	*zi-vu* doctor-NOM	*fu-jo* fairy-ACC	*gix* push	*fu-jo* fairy-ACC	*zi-vu* doctor-NOM	*gix* push
	(our SEGMENTED test type)					
Incorrect morphology	*zi-jo* doctor-ACC	*fu-vu* fairy-NOM	*gix* push	*fu-vu* fairy-NOM	*zi-jo* doctor-ACC	*gix* push
	(our UNSEGMENTED test type)					

3.5.2.3 Difficulty of the task

The third issue with the present study is that the task was simply too hard. This is most apparent from how many participants we had to exclude based on low accuracy: for Experiment 1, 101 participants out of the 183 we recruited; for Experiment 2, 43 out of 135 (this count is lower since we lifted the exclusion criterion for UNGRAMMATICAL sentences). A task this difficult would have encouraged participants to find the first reliable cue they could — the fixed word order — and cling onto it to succeed.

An obvious way to try to mitigate this issue would be to provide learners with more training. Based on Experiment 2 in Kenanidis et al. (2023), merely increasing the quantity of training trials may not substantially improve learning. However, in Rebuschat et al. (2021), participants learned to a relatively high degree an artificial language even more complex than ours — transitive sentences also included optional adjectives and

variable word order — from four training blocks of 48 trials each, interspersed with four testing blocks, over the course of about 45 minutes. This interleaving will benefit learning, as mentioned above.

An alternative way to make the experiment easier could be to use a different phenomenon than thematic role marking. We discuss another option, verbal tense morphology with tense adverbs, in Section 3.5.3.2 below.

3.5.3 Outlook and future directions

3.5.3.1 A follow-up using verbal production

As we have already acknowledged, our conception of a "production" task — asking participants to click on buttons to build up a sentence syllable-by-syllable — is not true language production. Specifically, our task clearly places lower retrieval and motor demands on participants than a verbal production task would have done. This observation may go some way to explaining why we have failed to replicate the benefits of language production that previous research describes. However, there is another crucial issue with our use of a text-based task: the training data already pushes participants in the direction of a word-level analysis, since in writing, the words are delimited by spaces, while the suffixes are not.

Both these issues point to the need for a follow-up study with no orthography i.e. using auditory stimuli in the training and judgement phases, and verbal production in the practice phase. Such a design would give learners the chance to segment the chunks for themselves, as they would do in real-world language learning (Havron & Arnon 2021; Siegelman & Arnon 2015), and to retrieve these chunks from memory as they practise the language, a process known to strengthen learning (Hopman & MacDonald 2018; Karpicke 2012; Karpicke & Roediger 2008; MacDonald 2013).

3.5.3.2 Testing a different phenomenon: Temporal reference

Perhaps the phenomenon of thematic role marking, with its requisite transitive sentences and large cast of characters, is too demanding for a brief experiment like this

one. A different linguistic phenomenon that can also be expressed both lexically and morphologically, and where the morphological cue is also redundant in the presence of the lexical one, is temporal reference (Sagarra & Ellis 2013).

For example, take a sentence like 'yesterday I walked'. It has a lexical cue to the past tense (the adverb *yesterday*) as well as a morphological cue (the suffix *-ed*). Sagarra and Ellis (2013) write that, for L2 learners of languages with verbal tense morphology like this, the adverbs tend to be acquired first, followed by the morphology.

Temporal reference might be a more viable phenomenon to test in an experiment like ours because we can instantiate it with syntactically simpler intransitive sentences that don't need to involve as many different characters. For example, we could envision an experiment in which participants learn sentences like "yesterday Annie walked" or "yesterday Annie swam". We would predict that participants who learn these sentences using a production task would be better at learning the verbal morphology than participants who learn the sentences using a comprehension task.

3.6 Conclusion

We began this investigation where two strands of previous research intersect, one showing that that language production helps learners identify and learn rules in their language (Hopman & MacDonald 2018; Izumi 2002; Swain 2005), and another showing that adults struggle to learn morphological rules and prefer word-level ones (Clahsen et al. 2010; DeKeyser 2005; Ellis 2022; Havron & Arnon 2021; Jordens et al. 1989; Lupyan & Dale 2010; Papadopoulou et al. 2011; Parodi et al. 2004; Sagarra & Ellis 2013). Bringing these observations together, we wanted to know whether a production task could help adult learners to identify a more difficult morphological rule over a more learnable word-level one.

Our results demonstrate that both L1 English and L1 German adults prefer to learn a word-level rule for marking thematic role over a morphological rule, even when they appear to notice morphological patterns — results that align with previous literature on how thematic role marking is learned (e.g. Grey et al. 2014; Kenanidis et al. 2023; Rebuschat et al. 2021). Contrary to our preregistered hypothesis, we didn't find that

practising a new language with a more active production-like task would steer learners away from this strong preference for word-level rules. But importantly, this does not mean that production has no effect — it just means that the current experiments don't provide support for or against the role of task effects for morphological rule learning. We still consider the question interesting and the hypotheses plausible, and we've proposed several avenues for future research to continue along the inroads we've made.

3.A Exploratory analysis: Removing the ungrammatical exclusion criterion

In the test phase of the experiment, we collected data that would inform two exclusion criteria: we would only keep participants who correctly accepted GRAMMATICAL sentences and correctly rejected UNGRAMMATICAL ones. We had defined "ungrammatical" as a word order that diverged from the one in training. Our reasoning was that participants should have learned that the language has SOV order, so they should reject the "ungrammatical" SVO order. Since SVO is also the basic word order of English, participants' rejection of it would provide the strongest test that they had learned the word order of the artificial language.

However, if a language has case marking, it is likely to also have free word order (Bentz & Winter 2013; Fedzechkina et al. 2011). It is therefore possible that participants who accepted sentences with a different word order had learned a case marking rule and associated that with a free word order, in which case our exclusion criterion would be removing exactly those participants who learned the segmented analysis we were targeting. This could explain why our results show such a strong preference for the unsegmented analysis.

Here, we lift this exclusion criterion and re-run the analyses described above. This criterion originally excluded 27 comprehension participants and 6 production participants; below we analyse data from 67 participants in the comprehension group and 46 in the production group.

3.A.1 Judgement

Figure 3.14 shows a similar pattern to Figure 3.3: a general preference for the novel sentences formed using the unsegmented analysis, and greater ambivalence toward ones formed with the segmented analysis.

We fit the same model described in Section 3.2.4 to this data. The pattern of results (shown in Table 3.8) remains the same as above. We conclude that the "ungrammatical" word order criterion did not exclude participants who learned a case marking rule

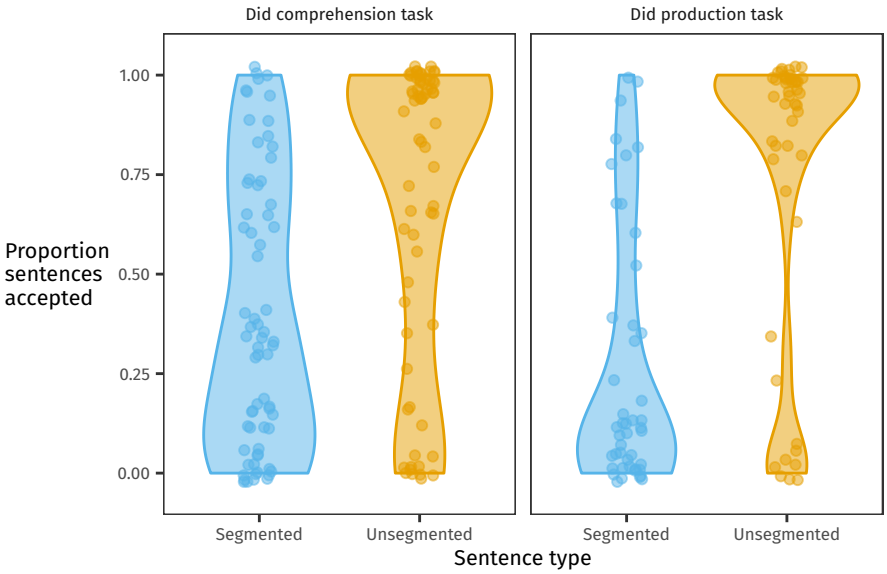

Figure 3.14: After lifting the ungrammaticality rejection criterion for participants in Experiment 1, the larger pool of participants show the same results: a strong preference for the unsegmented analysis over the segmented analysis, with no clear effect of task.

Table 3.8: Posterior distributions estimated by a model predicting sentence acceptance by condition, sentence type, and their interaction, now including data from participants originally excluded from Experiment 1 for rejecting sentences with a different word order than seen in training.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.41	0.24	–0.07	0.89
Condition	–0.19	0.47	–1.12	0.72
Sentence type	3.48	0.54	2.43	4.58
Condition:Sent. type	0.88	0.56	–0.20	1.98

and then extrapolated from it that word order was free.

3.A.2 Held-out character naming

The results from the held-out character naming analysis also remain extremely similar to the ones reported with the original exclusion criteria, as shown in Figure 3.15 and Table 3.9.

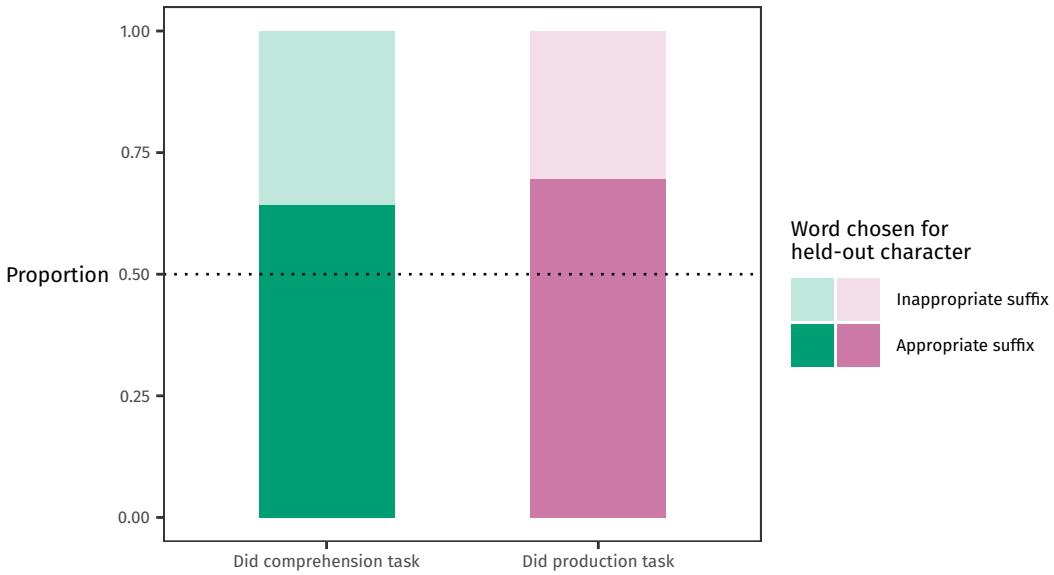

Figure 3.15: Proportion of Experiment 1 participants in each group, now including participants previously excluded from the analysis based on the ungrammaticality rejection criterion, who labelled the held-out character with a word containing the appropriate suffix. The same pattern holds as in the original analysis.

Table 3.9: Posterior distributions estimated by a model predicting appropriate suffix choice by condition, now including data from participants originally excluded from Experiment 1 for rejecting sentences with a different word order than seen in training.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.71	0.21	0.32	1.12
Condition	0.24	0.41	-0.55	1.06

3.B Overlaps in exclusion criteria

Table 3.10 shows how many of the 183 participants recruited for Experiment 1 were caught by each combination of exclusion criteria. (Gram. = incorrectly rejected GRAMMATICAL sentences; Ungram. = incorrectly accepted UNGRAMMATICAL sentences; Practice = low accuracy on practice phase; Notes = self-reported taking notes.)

Table 3.10: Full details of exclusions in Experiment 1

Gram.	Ungram.	Practice	Notes	Comprehension	Production
				40	40
			×	0	1
		×		5	12
	×			27	6
	×		×	1	0
	×	×		8	8
×				5	1
×		×		4	6
×	×			5	4
×	×	×		5	5

Table 3.11 shows how many of the 135 participants recruited for Experiment 2 were caught by each combination of exclusion criteria. (The ungrammatical sentences criterion was not used to exclude participants in Experiment 2.)

Table 3.11: Full details of exclusions in Experiment 2

Gram.	Ungram.	Practice	Notes	Comprehension	Production
				35	36
		×		4	10
	×			11	10
	×	×		2	1
×				4	2
×		×		1	4
×	×			4	2
×	×	×		7	2

3.C Analysis of all participants

In this appendix, we report the same analyses as in Sections 3.2.4 and 3.3.4 run on the data from all originally-recruited participants, imposing *none* of the preregistered criteria for exclusion.

3.C.1 Experiment 1

We recruited 183 participants in total for Experiment 1: 100 in the COMPREHENSION condition and 83 in the PRODUCTION condition.

3.C.1.1 Judgement

Figure 3.16 visualises the proportion of times each participant accepted each type of sentence at test. The same model described above was fit to this data; its posterior estimates are summarised in Table 3.12, and the conditional posterior distributions over the probability of accepting a sentence are shown in Figure 3.17.

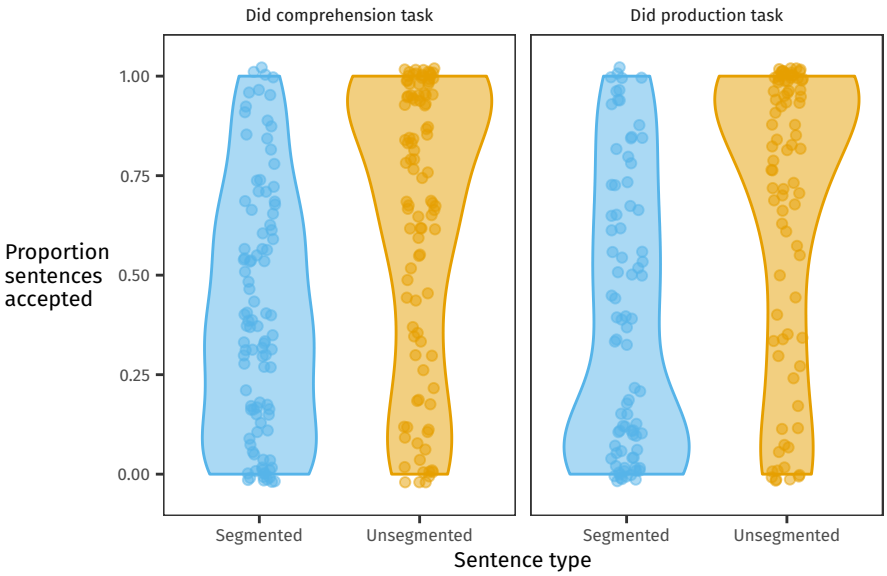

Figure 3.16: All 183 participants recruited for Experiment 1 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis, regardless of task. Each dot represents one participant's proportion of accepted sentences of each type.

Table 3.12: The posterior probability distributions estimated by the model for the sentence acceptance data from all 183 participants recruited for Experiment 1. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.27	0.16	–0.04	0.58
Condition	0.09	0.31	–0.54	0.69
Sentence type	2.07	0.32	1.45	2.71
Condition:Sent. type	0.23	0.32	–0.40	0.86

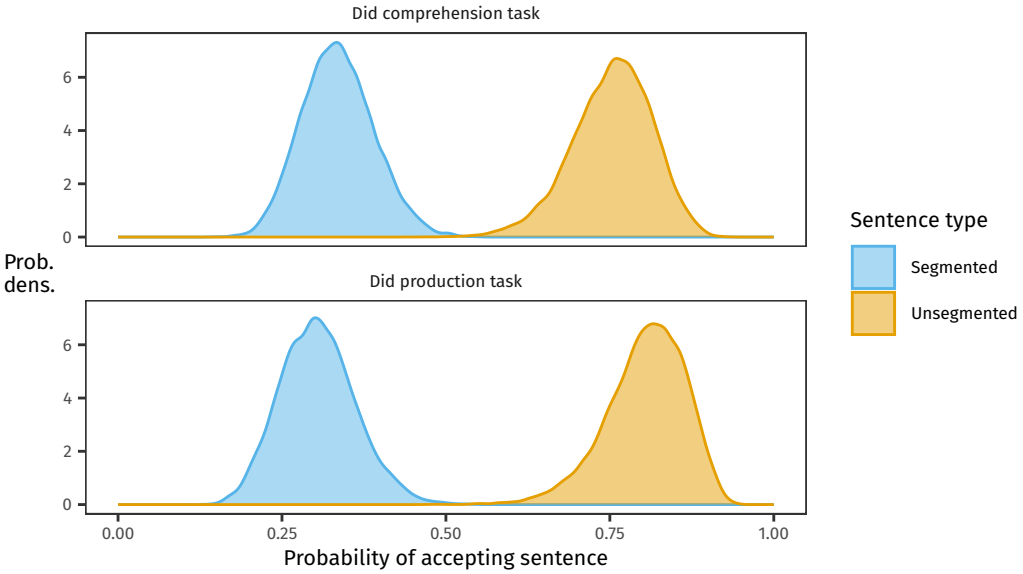

Figure 3.17: Conditional posterior probability distributions of the probability that all 183 participants recruited for Experiment 1 would accept a sentence. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

Overall, we see a similar pattern to the original analysis: participants in both the COMPREHENSION and the PRODUCTION condition accept the UNSEGMENTED sentences more than the SEGMENTED sentences.

3.C.1.2 Held-out character naming

Figure 3.18 illustrates the proportion of participants in each condition who named the held-out character using the appropriate suffix — the one that doesn't appear elsewhere in the sentence. As in the original analysis, more than half of the participants in both groups chose the word containing the appropriate suffix, and the model estimates that both groups have very similar probabilities of selecting the appropriate suffix (see the posterior summaries in Table 3.13 and the conditional posterior distributions in Figure 3.19).

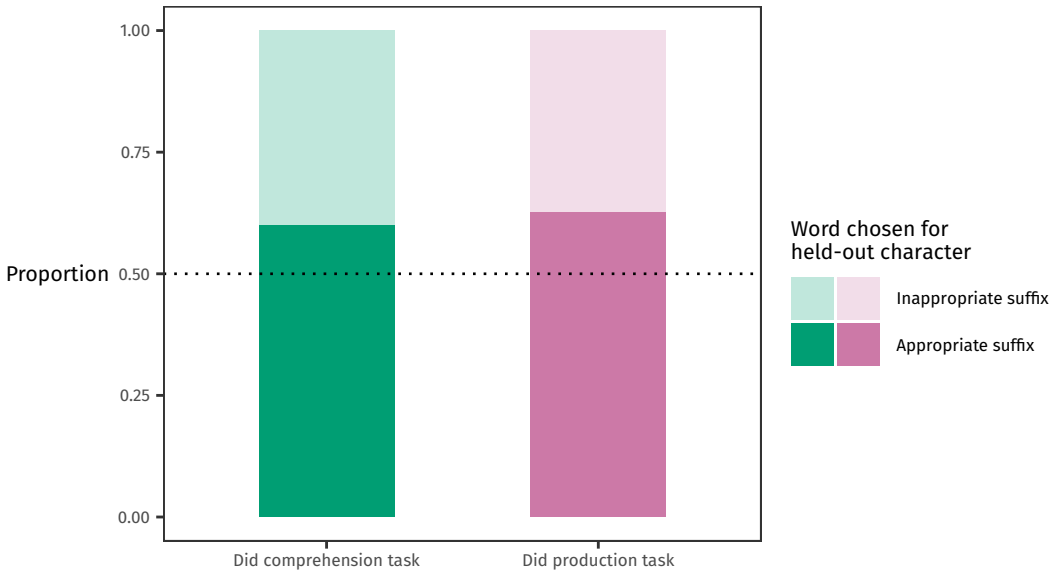

Figure 3.18: In the held-out character naming task of Experiment 1, more than half of all 183 participants selected the word with the appropriate suffix.

Table 3.13: The posterior probability distributions estimated by the model for all 183 participants' held-out character naming data in Experiment 1. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.46	0.15	0.17	0.76
Condition	0.11	0.31	–0.48	0.72

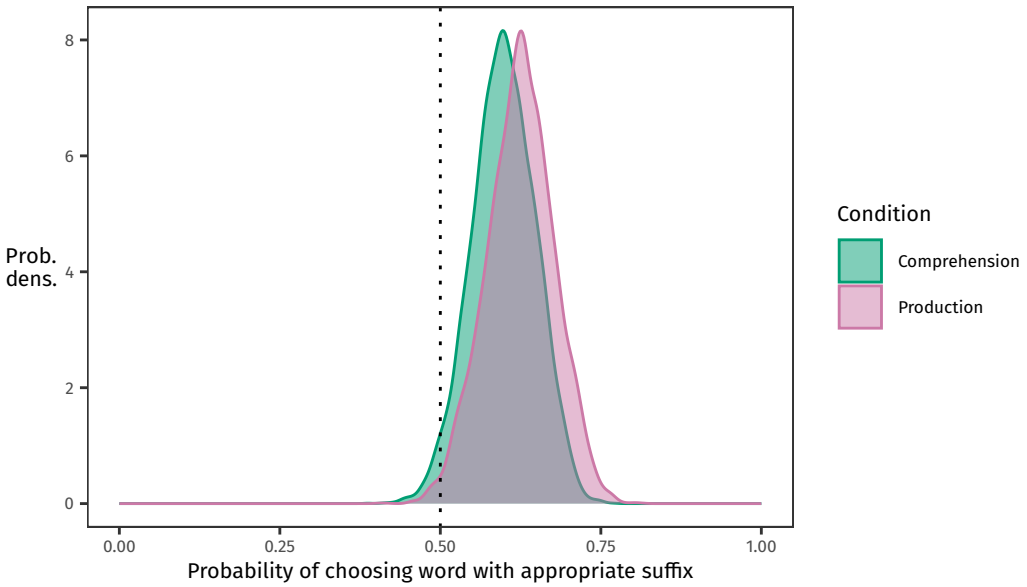

Figure 3.19: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 1, shown for all 183 originally recruited participants.

3.C.2 Experiment 2

We recruited 135 participants in total for Experiment 2: 68 in the COMPREHENSION condition and 67 in the PRODUCTION condition.

3.C.2.1 Judgement

In Figure 3.20, we show the proportion of times each participant accepted each type of sentence at test. We see the same pattern as in the original Experiment 2 data and in the data of all 183 participants from Experiment 1: participants prefer the UNSEGMENTED sentences over the SEGMENTED ones, regardless of task. Table 3.14 summarises the posterior distributions estimated by the same model as above, and Figure 3.21 shows the conditional posterior distributions.

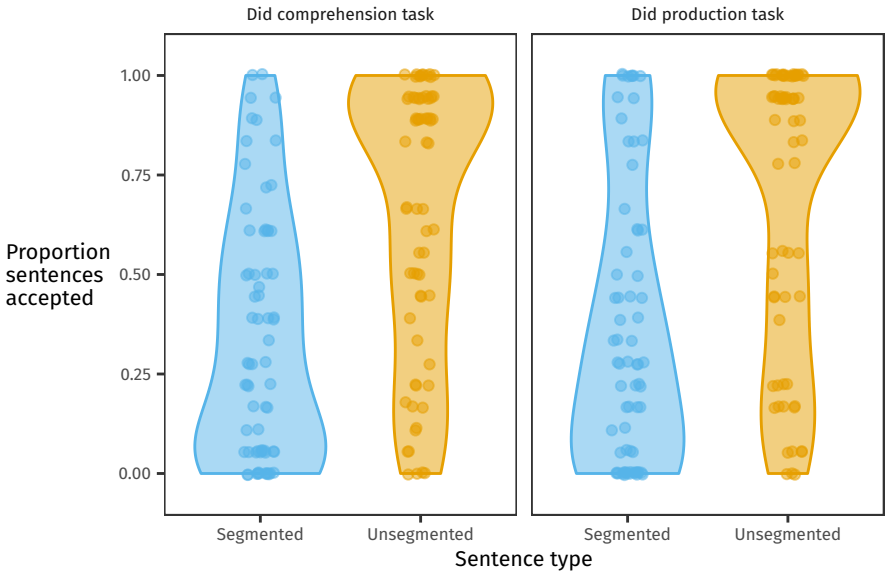

Figure 3.20: All 135 participants recruited for Experiment 2 accepted novel sentences that followed the unsegmented analysis more frequently than sentences that followed the segmented analysis, regardless of task. Each dot represents one participant's proportion of accepted sentences of each type.

Table 3.14: The posterior probability distributions estimated by the model for the sentence acceptance data from all 135 participants recruited for Experiment 2. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	0.15	0.14	-0.13	0.43
Condition	0.32	0.27	-0.21	0.87
Sentence type	2.51	0.41	1.72	3.33
Condition:Sent. type	0.19	0.41	-0.61	0.98

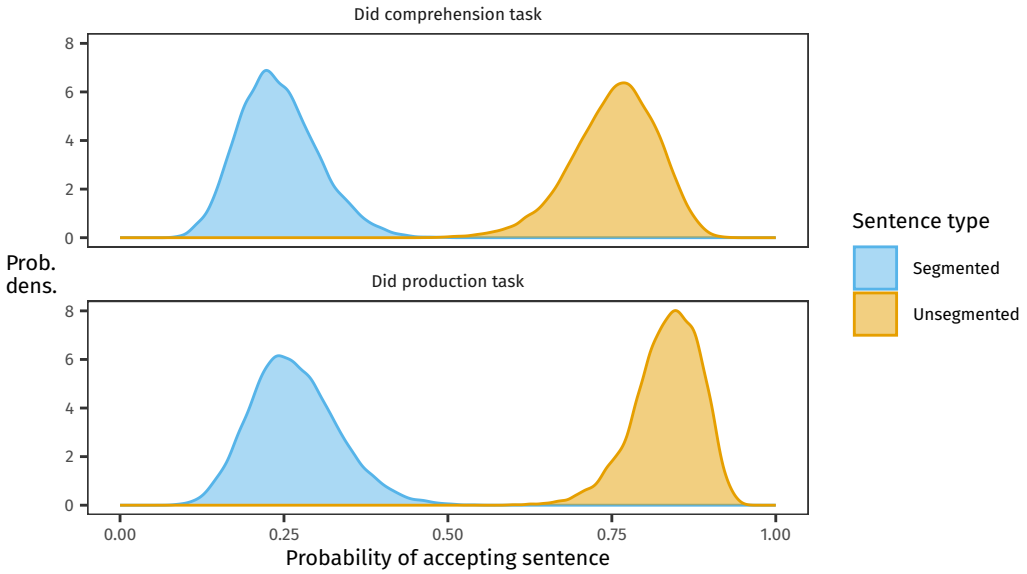

Figure 3.21: Conditional posterior probability distributions of the probability that all 135 participants recruited for Experiment 2 would accept a sentence. UNSEGMENTED sentences are more likely to be accepted than SEGMENTED sentences, regardless of whether participants did a comprehension or production task.

3.C.2.2 Held-out character naming

Figure 3.22 shows that, like the original analysis, around three-quarters of participants in each condition named the held-out character using the appropriate suffix. Table 3.15 summarises the posteriors estimated by the same model described above, and Figure 3.23 shows the conditional posterior distributions.

Table 3.15: The posterior probability distributions estimated by the model for all 135 participants' held-out character naming data in Experiment 2. Values are on the log-odds scale.

	Estimate	Est'd error	Lower 95% CrI	Upper 95% CrI
Intercept	1.26	0.21	0.86	1.67
Condition	0.14	0.41	–0.65	0.95

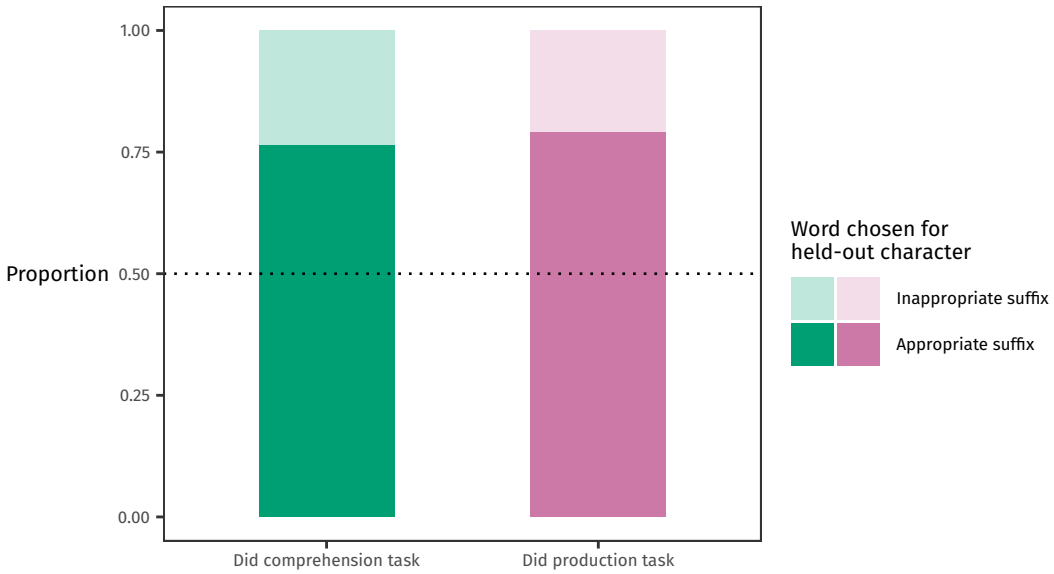

Figure 3.22: In the held-out character naming task of Experiment 2, at least three-quarters of all 135 participants selected the word with the appropriate suffix.

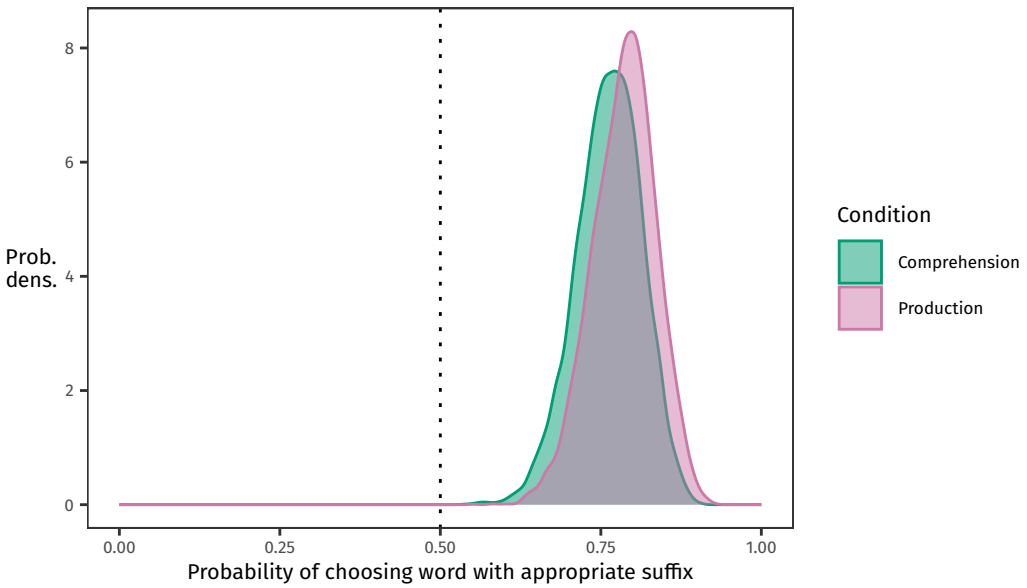

Figure 3.23: Conditional posterior probability distributions over the probability of selecting a word that contains the appropriate suffix in Experiment 2, shown for all 135 originally recruited participants.

3.D Bayesian model specifications

Details of the prior predictive checks used to arrive at these prior specifications are available in Appendix A of Elizabeth's thesis (Pankratz 2025).

3.D.1 Judgement

```
brm(
  sentence_accepted ~ sent + cond + sentcond + (sent | ppt_id),
  family = bernoulli(),
  prior = c(
    prior(normal(0, 1.5), class = Intercept),
    prior(normal(0, 2), class = b),
    prior(normal(0, 5), class = sd, coef = Intercept, group = ppt_id),
    prior(normal(0, 5), class = sd, coef = sent, group = ppt_id),
    prior(lkj(2), class = cor, group = ppt_id)
  )
)
```

3.D.2 Held-out character naming

```
brm(
  match ~ cond,
  family = bernoulli(),
  prior = c(
    prior(normal(0, 1.5), class = Intercept),
    prior(normal(0, 2), class = b)
  )
)
```

3.E Exploratory analysis: Participants who know case marking languages

In the post-experiment debrief questionnaire for Experiment 1, we asked participants if they knew or understood any other languages beyond English. If they self-reported knowing a case marking language, we placed them into a separate group from the participants who did not. Fifteen participants out of 80 reported that they know or understand the following case marking languages: Arabic, Czech, German, Latin, Polish, Romanian, Slav, Somali, Tunisian, Turkish, and Urdu.

3.E.1 Judgement

Figure 3.24 visualises the proportion of sentence acceptance judgements for each participant, split by condition and further by whether each participant knows a case marking language.

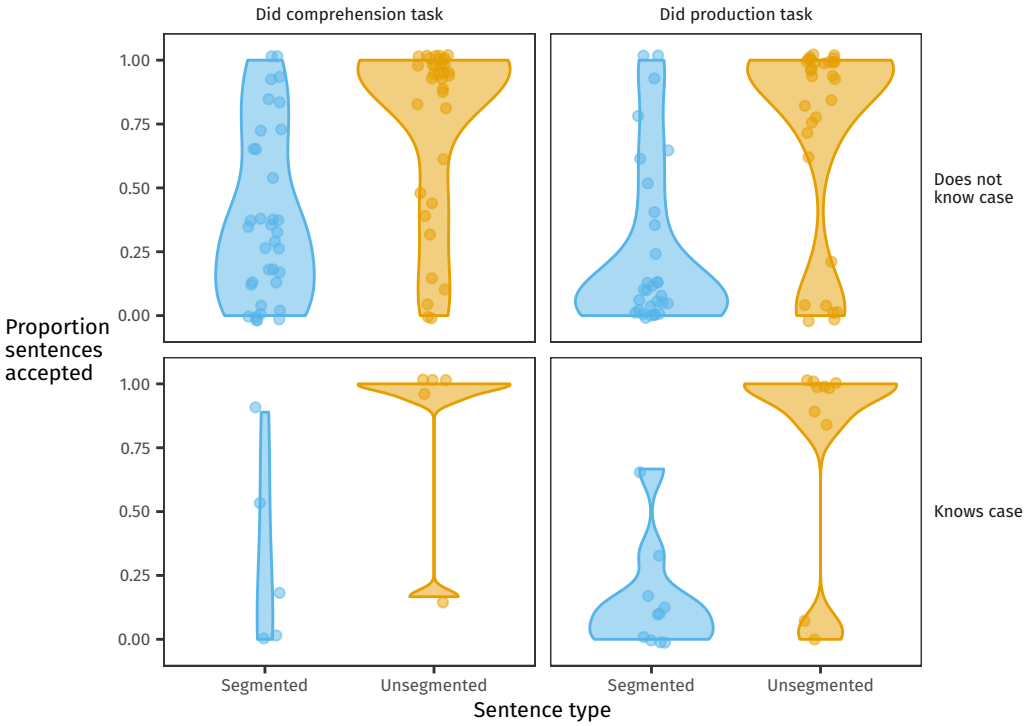

Figure 3.24: Participants in Experiment 1 who self-reported knowing a case marking language show a similar pattern of sentence acceptance to participants who do not know a language with case marking.

We fit the same Bayesian model as described in Section 3.2.4 to this data, adding in an additional sum-coded predictor for knowledge of case (–0.5 when the participant does not know a case marking language, +0.5 when they do), and all two- and three-way interactions with the predictors sentence type and condition (scaled to ±0.5). Table 3.16 summarises the posterior distributions of the population-level effects estimated by the model. In short, the previously-estimated effects remain qualitatively the same, and the model indicates great uncertainty about any association between prior knowledge of case marking languages and acceptance of sentences formed using the segmented analysis.

Table 3.16: Posterior distributions estimated by a model predicting sentence acceptance by condition, sentence type, and knowledge of a case marking language, and all interactions between them.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.53	0.35	–0.15	1.25
Condition	–0.76	0.66	–2.05	0.52
Sentence type	4.36	0.86	2.68	6.03
Case	–0.03	0.69	–1.36	1.37
Condition:Sent. type	0.52	0.84	–1.11	2.18
Condition:Case	0.09	0.67	–1.21	1.42
Sent. type:Case	0.44	0.83	–1.21	2.04
Cond.:Sent. type:Case	0.28	0.85	–1.38	1.93

3.E.2 Held-out character naming

Figure 3.25 illustrates that the 15 participants who know a case marking language select the word with the appropriate suffix less often than the larger group of 65 participants who do not know case. However, we fit a model estimating appropriate suffix choice as a function of condition, knowledge of case, and their interaction (scaled to ±0.5), and the posterior distribution estimates in Table 3.17 indicate that we cannot be certain about any differences between participant groups.

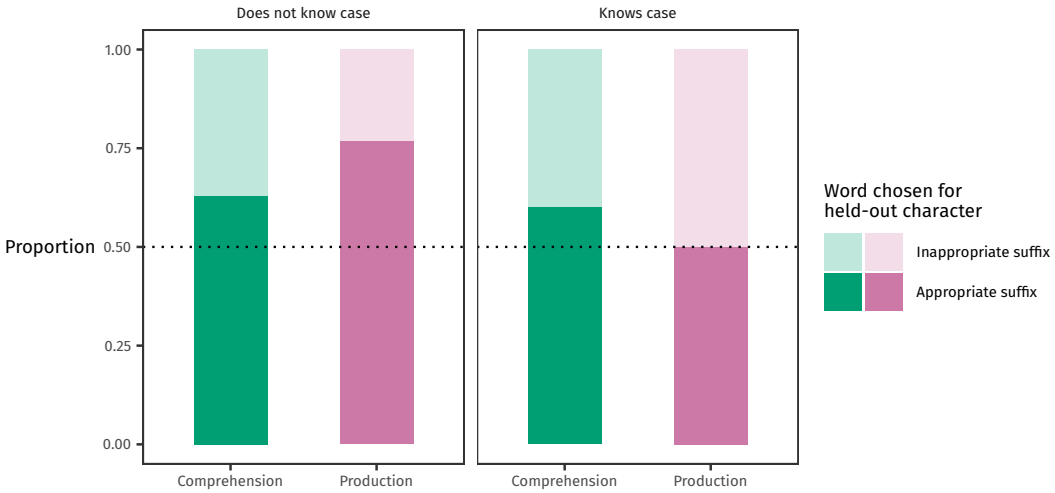

Figure 3.25: The 15 participants in Experiment 1 who know a case marking language give overall less appropriate responses to the held-out character naming task, with production participants selecting the appropriate suffix less than participants in the comprehension group.

Table 3.17: Posterior distributions estimated by a model predicting appropriate suffix choice by condition, knowledge of a case marking language, and their interaction.

	Estimate	Est.Error	Q2.5	Q97.5
Intercept	0.56	0.32	–0.06	1.19
Condition	0.10	0.62	–1.11	1.27
Case	–0.59	0.62	–1.82	0.65
Condition:Case	–0.56	0.62	–1.78	0.64

Chapter 4

The evolution of phonetic clustering in the lexicon

I'm sorry, but if the other player was real,
they weren't using their brain properly.

Anonymous Prolific participant

Author contributions

The first element of this chapter is a small corpus study. I was assisted in gathering and preparing the data I used for the first iteration of this study by a fellow PhD student, Juan Guerrero Montero (although the version reported in this chapter uses the data from Dautriche et al. (2017a, 2017b), available at https://osf.io/rvg8d/). Juan also provided the code to generate artificial words that conformed to English phonotactics; a full list of the rules included in his code is provided in Appendix 4.F. I came up with the research question, developed the methodology, conducted the analysis, and wrote up the results.

The second element of this chapter (including Appendix 4.A) is a reproduction of a paper submitted to *Cognition* in December 2024 and published on *PsyArXiv* as a preprint. The paper is under review at the time of writing. Some minor details differ between the version contained in this thesis and the version published on *PsyArXiv*,

and there may be more major discrepancies in any future published version; this is due to divergences between corrections received on my thesis and comments received during the peer review process. The paper was co-authored with my two supervisors, Jennifer Culbertson and Simon Kirby. I developed the model described in Section 4.3 independently. The design for the experiments reported in Section 4.4 and Appendix 4.A was developed during supervision meetings where all three authors were present and contributing. I created the experiment software, collected the data, conducted the analysis, and wrote the first draft of the paper. Both co-authors provided feedback during the writing and revision of the paper.

After Appendix 4.A, all other appendices appear only in this thesis. Appendix 4.B contains pilot data from a follow-up experiment using an oral production testing task; the data for this experiment was collected by an undergraduate research assistant, Beth Kipling, under my supervision. The work presented in Appendices 4.C through 4.E was conducted independently.

Open materials

All materials, code and data used for this chapter are freely available at https://osf.io/vsy6z/.

4.1 How similar are words? A corpus study

This chapter started from a different question than the one I ended up answering. In the first instance, I was interested in Zipf's Law of Abbreviation, and wanted to know whether longer words were more distinctive than shorter words only to the extent predicted by their length, or above and beyond this level. I started playing with some data from the Google Books corpus, and quickly realised that something strange was going on: in English at least, words of *all lengths* were more similar to each other than they really needed to be. Unfortunately though, as I soon found out, this was not a new discovery: Dautriche et al. (2017a) had already published an extensive analysis of four languages (English, German, Dutch and French) which came to the same conclusion — that lexicons are surprisingly *phonetically clustered*.

Although our results were very similar, I did take a slightly different approach to determining what it meant for a lexicon to be "surprisingly" clustered, so I just want to briefly lay out that methodology here and show that it replicates Dautriche et al.'s key findings.

4.1.1 Data

For ease of comparison, the data I use here is the same as that used by Dautriche et al. (2017a, 2017b), which I downloaded from the OSF repository associated with the latter: https://osf.io/rvg8d/. This dataset contains 6,197 words of English tagged by CELEX (Baayen et al. 1995) as monomorphemic. Words are phonemically transcribed (with diphthongs transformed into 2-character strings) and homophones are discarded. The distribution of word lengths (in number of phonemes) is given in Figure 4.1.

It's worth saying that I obtained the same pattern of results from this data and in my original exploratory study using the Google Books corpus. However, the CELEX data is more conservative, since it is morphologically parsed; in the Google Books data, I was not able to control for productive morphology as a source of phonetic clustering.

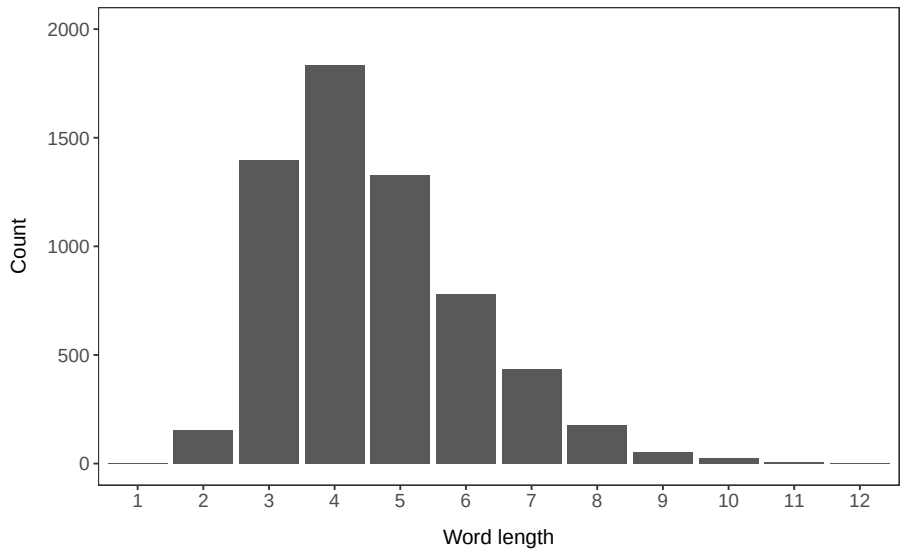

Figure 4.1: Distribution of word lengths (in number of phonemes) for the 6,197 English words tagged by CELEX (Baayen et al. 1995) as monomorphemic.

4.1.2 Simulated baselines

The core of my approach is the same as that used by Dautriche et al. (2017a): benchmark a real language against a simulated baseline constructed from that language. I evaluated the clustering statistics of the real lexicon against six different baselines, ranging from very random to very phonotactically constrained. The random baselines were included just as a proof of concept: I do not consider them to be plausible models of what English could actually look like. For each baseline, I generated the same number of unique words of each length as in the real English lexicon.

4.1.2.1 Random baselines

Random strings The most basic model imaginable is one in which we just randomly concatenate English phonemes until we reach the desired word length. More specifically, each word in this model is generated by repeatedly sampling with replacement from the set of phonemes present in the real lexicon.

By-word shuffle Potentially a slightly less random model is one in which new words are generated from existing ones. Concretely, for each of the 6,197 words in the real lexicon, I create a corresponding baseline word by randomly shuffling the phonemes of the real word.

By-position shuffle An even less random model is one which pays some attention to what position certain phonemes tend to occupy within words e.g. the fact that [ŋ] cannot occur word-initially in English. To achieve this, I performed a by-position shuffle of the real words. To understand how this works, imagine that each real word occupies a row of a dataframe, and the columns correspond to individual phonemes: Column 1 contains each word's first phoneme, Column 2 the second phoneme, and so on. To generate a set of new words, I performed a column-wise shuffle of the real words of each length. In other words, each new word is created by sampling (without replacement), for each position, one phoneme from the list of phonemes which appeared in that position in the real words of the target length. For example, given a toy lexicon {dog, pig, cat}, this procedure could generate a simulated lexicon {pot, cag, dig}: there is still one word beginning with each of {c, d, p}, one word with each of {a, i, o} in the middle, two words ending with g and the other with t.

4.1.2.2 Constrained baselines

Dautriche et al. (2017a) used a 5-phone model to create their simulated lexicons. In this model, words are generated phoneme-by-phoneme, with the probability of choosing a particular phoneme in any position proportional to its probability of following the previous 4 in the real lexicon. Even before seeing their paper, I was concerned that n-phone models in general would significantly over-fit to the real data; indeed, more than half of the words generated by Dautriche et al.'s 5-phone model were actual words of English. For my own n-phone models, I therefore focussed on a smaller sequence length (trigrams), which I felt would be constrained enough to generate mostly phonotactically legal sequences, but free enough to generate a higher proportion of plausible but unattested words. I also implemented a phonotactic model that paid no attention to frequencies in the corpus data, instead sampling uniformly from sequences deemed legal under a set of hard constraints on syllable structure.

Before setting out the details of these baselines, I want to make a brief detour to discuss the relationship between phonotactics and clustering. As in Dautriche et al. (2017a), the intention of these baselines is to provide a null hypothesis for how clustered or disperse lexicons would be as a result of phonotactics alone. Clearly, phono-

tactics is itself a source of clustering, because it constrains the space of possible wordforms in a language. However, if real lexicons turn out to be even *more* clustered than phonotactically-controlled baselines, this may suggest that they are being shaped by a pressure in favour of clustering. It should be noted, though, that while many phonotactic constraints are language-specific (e.g. words can't start with *zb* in English, but they can in Polish), this does not necessarily mean that phonotactics is entirely arbitrary. In particular, there are many well-known typological tendencies in phonology — like vowel harmony (Finley & Badecker 2008, 2010; A. Martin & Peperkamp 2020; A. Martin & White 2021), or the Sonority Sequencing Principle (Clements 1990; Greenberg 1965; Zec 1995) — which are argued to be the result of adaptations to articulatory *naturalness*, or substantive bias (e.g. Archangeli & Pulleybank 1994; Blevins 2004; Donegan & Stampe 2009; Finley & Badecker 2007; Hayes & White 2013; Moreton & Pater 2012; Zheng & Do 2025). Some of these adaptations may naturally result in higher levels of clustering for reasons independent of a clustering pressure. For example, in the case of vowel harmony, if two words happen to share the first vowel, they're also more likely to share the same vowel in subsequent syllables; however, this resemblance between words arose not from a pressure to maximise similarity across the lexicon as a whole, but from a pressure to maximise similarity *within* individual words. It is also entirely possible that some more arbitrary-seeming phonotactic constraints could arise precisely from the clustering pressure we are trying to detect i.e. languages might preferentially impose constraints that result in a more restricted space of possible wordforms (over those that allow for greater dispersion in the lexicon). This possibility is, however, very speculative, and does not seem to have received any attention in the literature; if anything, the discussion tends to focus in the opposite direction, on the role of perceptual salience and the need for phonotactics to allow for sufficient *distinctiveness* of wordforms (e.g. Baroni 2014; Dziubalska-Kołaczyk 2014; Liljencrants & Lindblom 1972). In any case, the point of all this is just to highlight that, even if languages don't end up looking significantly more clustered than their baselines, this does not provide evidence *against* a clustering pressure: it may simply be that phonotactic constraints themselves — potentially for a mix of reasons — already give rise to such high levels of clustering that no effects are observable above and beyond this.

Attested trigrams The most unrestricted of my phonotactically constrained baselines generates new words by concatenating phonemes from the real lexicon in such a way that all trigrams (sequences of 3 phonemes) are attested in English. For example, [spæŋ] would be a possible word under this model since all trigrams are attested in English: [►►s], [►sp], [spæ], [pæŋ] (where ► represents start-of-word symbols). I generated each word phoneme-by-phoneme, each time choosing the next phoneme by randomly sampling from those that had non-zero probability of following the previous two in the English words of the target length.

Probabilistic phonotactics This model is very similar to the previous one, but more closely fitted to the probabilistic patterns in the English data. Specifically, when generating a word, I sample each phoneme according to its *probability* of following the previous two in the English words of the target length, rather than by sampling uniformly from those with non-zero probability. This again has the effect of creating words composed only of attested trigrams, but trigrams that are more frequent in the English data will also be more frequent in the simulated data. Unlike in the previous trigram model, here I also used Laplace smoothing with parameter 0.01 to assign non-zero probability to unseen trigrams.

Rule-based phonotactics Finally, to avoid over-fitting to the English data, I wanted to look at a baseline where words were generated to conform to the phonotactic rules of English, but with no heed to the frequency of particular sound sequences. A set of rules was extracted from the Longman Pronunciation Dictionary of American English (Wells & Hung 1990); a full list is available in Appendix 4.F. Each word is created by generating syllables until the target word length is reached. Each *syllable* is generated by sampling an onset, nucleus and coda, and checking that they are compatible with each other. Finally, once all the syllables are sampled, they are checked for compatibility with regard to intersyllabic restrictions and restrictions at the end of the word.

This is arguably the most interesting model of what English *could* look like, since it generates words that satisfy all hard phonotactic constraints and avoids overfitting to patterns that may arise because of, for example, historical relatedness. However, it should be noted that it does not take account of violable but robust probabilistic ten-

dencies, such as similar place avoidance and the obligatory contour principle (Cathcart 2024; Frisch et al. 2004; Pozdniakov & Segerer 2007). Potentially a better model to test in future work would be one that captures both categorical and gradient phonotactic patterns (e.g. Hayes & Wilson 2008).

4.1.3 Analysis

For each word length from 3 to 8 (inclusive), I calculated three measures of clustering for each lexicon — real and simulated.

Neighbourhood density is probably the most commonly used measure in the literature, and quantifies, for a given word, how many other words can be created by making a single edit (insertion, deletion, or substitution). Note that, since I only compare words of a given length to other words of that length, substitution is the only relevant operation; this means that this measure could equally be described as a count of **minimal pairs** (following the terminology used by Dautriche et al.). If the real lexicon is more clustered than expected by chance, this would be reflected in *higher* average neighbourhood density (relative to the baselines).

Clustering coefficient (C) is a measure from graph theory, which here captures the extent to which a word's neighbours are also neighbours of each other. Given a word i in an undirected graph, C_i is given by:

$$C_i = \frac{2 \cdot |\{e_{jk}\}|}{k_i \cdot (k_i - 1)}$$

where e_{jk} refers to the presence of a connection between two neighbours (j and k) of word i, $|...|$ indicates cardinality (the number of elements in the set), and k is the degree (i.e. neighbourhood density) of word i (Vitevitch et al. 2011). If the real lexicon is more clustered than expected by chance, this would be reflected in *higher* average clustering coefficient (relative to the baselines).

Finally, **Levenshtein distance** is a measure of string edit distance, which quantifies the number of insertions, deletions and substitutions required to get from one string to another. If the real lexicon is more clustered than expected by chance, this would be

reflected in *lower* average Levenshtein distance (relative to the baselines).

4.1.4 Results and discussion

Results for **neighbourhood density** are shown in Figure 4.2. Unsurprisingly, neighbourhood density is considerably lower for the three random baselines than the real lexicon, since these baselines have no or few restrictions on which phonemes can occur in which positions, thus allowing words to differ from each other a lot more. However, this same pattern also holds for the constrained baselines, with the exception of words of length 3 generated using the attested trigram model (for reasons that are unclear to me). In other words, real words tend to have more minimal pairs than words generated under plausible models of English phonotactics.

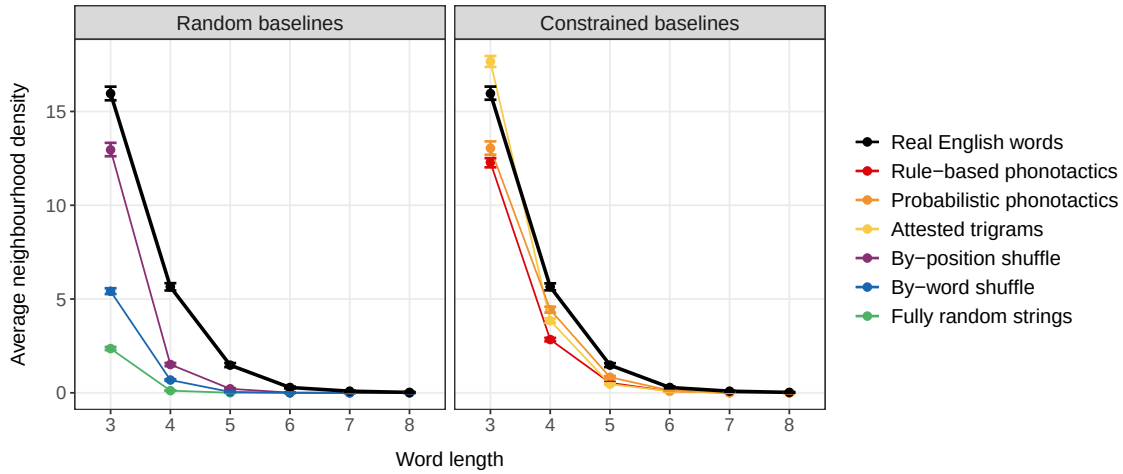

Figure 4.2: Average neighbourhood density for real English words compared to the six baselines. Points represent means; error bars represent 95% confidence intervals. Unsurprisingly, neighbourhood density is considerably lower for the three random baselines than the real lexicon. However, this same pattern also holds for the constrained baselines, with the exception of words of length 3 generated using the attested trigram model.

Results for **clustering coefficient** are shown in Figure 4.3. Again, unsurprisingly, clustering coefficient is considerably lower for the three random baselines than the real lexicon. However, the same general pattern also holds for the constrained baselines, particularly for longer words and under the rule-based phonotactics model. For shorter words, both trigram models look very similar to the real data. Overall though, it seems that in addition to real words having more neighbours, they might also come from neighbourhoods that are more well-connected than the neighbourhoods in the simulated lexicons.

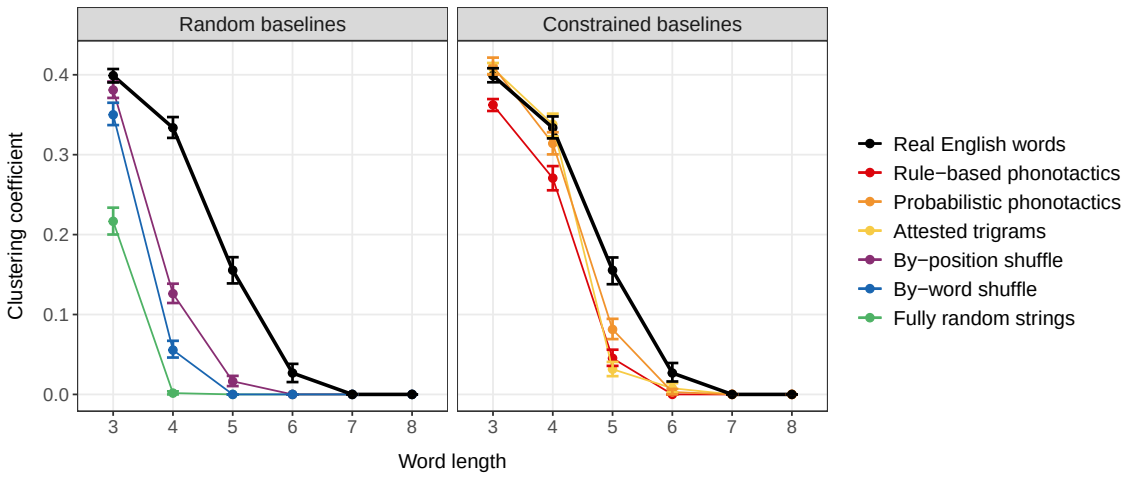

Figure 4.3: Average clustering coefficient for real English words compared to the six baselines. Points represent means; error bars represent 95% confidence intervals. Unsurprisingly, clustering coefficient is considerably lower for the three random baselines than the real lexicon. However, the same general pattern also holds for the constrained baselines, particularly for longer words and under the rule-based phonotactics model.

Finally, results for **Levenshtein distance** are shown in Figure 4.4. The differences here are (numerically) much smaller than for the other two measures, since the Levenshtein distance between two words is largely a product of their lengths. However, the difference becomes more pronounced for longer words, and again, it seems that real words are more similar to each other than those of the simulated lexicons — random and constrained. The only notable exception to this trend is the by-position shuffle, which closely approximates the real data for all word lengths[1].

[1]This is as expected: this method of generating words maintains the position of each phoneme within the set of words, so average edit distance is largely unaffected. To illustrate, consider again the example toy lexicon I described above {dog, pig, cat}, and the simulated lexicon created with the by-position shuffle {pot, cag, dig}. In the real lexicon, edit distance between "dog" and "pig" is 2, between "dog" and "cat" is 3, and between "pig" and "cat" is 3 (average = 2.33). In the simulated lexicon, edit distance between "pot" and "cag" is 3, between "pot" and "dig" is 3, and between "cag" and "dig" is 2 (average = 2.33).

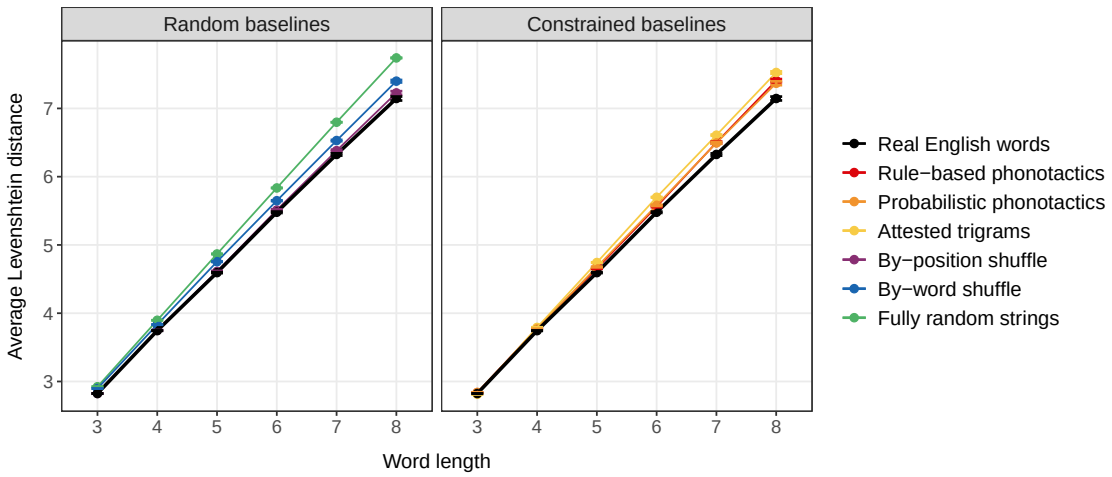

Figure 4.4: Average Levenshtein distance for real English words compared to the six baselines. Points represent means; error bars represent 95% confidence intervals. Although Levenshtein distance is primarily a function of word length in both the real and the simulated lexicons, a small difference is visible for longer words, whereby the real English words are more similar to each other than those in all baselines (with the exception of the by-position shuffle).

On the whole, these results suggest (in line with those of Dautriche et al. 2017a) that the lexicon of English is surprisingly phonetically clustered: words are more similar to each other than they really need to be given the available phonotactics. Dautriche et al. (2017a) also showed that this pattern generalises to other languages. In the paper reproduced in the next section, I turn to the logical follow-up question: *why* are lexicons organised in this way?

The lexicon adapts to competing communicative pressures: Explaining patterns of word similarity

Abstract

Cross-linguistically, lexicons tend to be more phonetically clustered than required by the phonotactics of the language; that is, words within a language are more similar to each other than they need to be. In this study, we investigate how this property evolves under the influence of competing communicative pressures: a production-side pressure to re-use more easily articulated sounds, and a comprehension-side pressure for distinctiveness of wordforms. In an exemplar-based computational model and a communication experiment using a miniature artificial language, we show that natural-language-like levels of clustering emerge from a trade-off between these pressures. With only one pressure at work, the resulting lexicons tend to inhabit an extreme region of the possible design space: production pressures alone give rise to maximally clustered lexicons, while comprehension pressures alone give rise to maximally disperse lexicons. We also test whether clustering emerges more strongly for high-frequency items, but our results lend support only to a weak relationship between frequency and clustering. Overall, this study adds to a growing body of evidence showing that mechanisms operating at the level of individual language users and individual episodes of communication can give rise to emergent structural properties of language.

4.2 Introduction

Different languages have different rules about how sounds can be combined to form words. For example, "zad" is an unattested but possible word of English, whereas "zbad" is both unattested and impossible (but could be a word of Polish). Naturally, the fact that these rules differ between languages means that words within a language generally sound more similar to each other than they do to words of other languages. Indeed, both infants (Jusczyk et al. 1993; Mehler et al. 1988; Moon et al. 1993) and adults (Lorch & Meara 1989; Marks et al. 2003; Stockmal et al. 1996) can discriminate surprisingly well between languages, even ones they don't know.

Perhaps less obvious is the fact that, even within a language, possible sounds and sound combinations are not necessarily equally frequent. Figure 4.5 gives a sense that, while "zad" is a phonotactically legal sound sequence in English, it is perhaps not very likely to be coined as a new word: the [z] phoneme is relatively uncommon in English (especially in word-initial position), and the [zæ] biphone is extremely low-frequency. This skewed distribution is not unique to English: it is a common property across languages that not all possible sounds or sound sequences are equally frequent (Krevitt & Griffith 1972; Macklin-Cordes & Round 2020; Martindale et al. 1996). As a result, words within a language are actually more similar *to each other* than they really need to be. In other words, lexicons are *phonetically clustered*.

Naively, we might expect languages to use up their available phonotactic space more uniformly; that is, words could be evenly distributed in this space to avoid repeating sound sequences where possible. Successful communication depends on listeners being able to perceive and interpret a speaker's message with a high degree of accuracy. And since communication takes place over a noisy channel (Gibson et al. 2013a; Levy 2008; Shannon 1948), there is always a possibility that information will be lost; a lexicon that maximised the distance between words would reduce this possibility (Flemming 2004). Indeed, we know that comprehension is easier when words are more distinct: in line with the Neighbourhood Activation Model (Luce & Pisoni 1998), words from sparser phonological neighbourhoods and less densely connected areas of the lexical network (i.e. words that are less similar to other words) are recognised more quickly and accurately, especially in noisy conditions (Chan & Vitevitch 2009; Cluff &

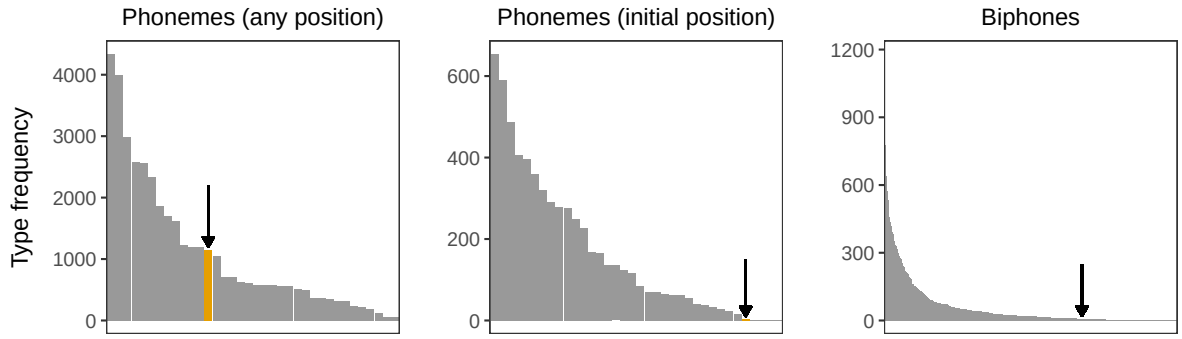

Figure 4.5: Type frequency of all phonemes and biphones of English, derived from the British National Corpus (BNC Consortium 2007) using List 1.2 (rank frequency list for the whole corpus, limited to words with a frequency of at least 100 per million) from Leech et al. (2001), converted to IPA using the `eng-to-ipa` package in Python (https://pypi.org/project/eng-to-ipa/). Yellow bars and arrows indicate the [z] phoneme in the left-hand and middle panels, and the [zæ] biphone in the right-hand panel. The specific identity of other phonemes/biphones is not shown on the x-axis for ease of presentation; there are 36 unique phonemes and 670 unique biphones represented in the word list. The key observation is that the shape of all these distributions is skewed: certain sounds and sound sequences are considerably more frequent than others.

Luce 1990; Goldinger et al. 1989; Magnuson et al. 2007; Siew & Vitevitch 2016; Vitevitch & Luce 1998).

However, the effect of word similarity on comprehension is not completely straightforward. In particular, increases in phonotactic probability (which reflects the existence of high-frequency sound sequences within a word) have been found to be beneficial for word recognition (Vitevitch & Luce 1998, 1999; Vitevitch et al. 1997, 1999) Furthermore, there is good evidence that spoken word *production* is facilitated by increases in both neighbourhood density *and* phonotactic probability (Chen & Mirman 2012; Gahl et al. 2012; Goldrick & Larson 2008; Goldrick & Rapp 2007; Munson 2001; Stemberger 2004; Vitevitch & Luce 1998, 2005; Vitevitch & Sommers 2003; Vitevitch et al. 2004). That is, words that are more similar to other words are generally pronounced more quickly and accurately.

This suggests that communication involves a complex interplay of different functional pressures coming from both production and perception, and taken together these do not straightforwardly point to an overall advantage or disadvantage of word similarity. How might language users balance these competing pressures in a way that leads to phonetically clustered lexicons? Almost 80 years ago, the linguist George Kingsley Zipf claimed that the organisational structure of languages is shaped by a

trade-off between a pressure for accurate communication on the one hand, and a pressure for efficiency on the other (Zipf 1949). Although this claim is most famously instantiated in the "Law of Abbreviation" — whereby more frequent words tend to be shorter — Zipf also argued that languages should preferentially re-use easy-to-articulate sounds over more difficult sounds (Zipf 1935). A related argument was made by Piantadosi et al. (2012), who suggest that an efficient communication system should re-use more easily produced words and sounds, even if doing so results in some ambiguity.

Of course, there are several reasons why lexicons might re-use particular sounds more than others (as in Figure 4.5), not all of which point to an adaptive explanation. For example, we would expect certain sounds to reoccur across many words in languages with productive morphology: *unkind*, *unsatisfying* and *unpleasant* all sound somewhat similar because of a shared prefix, while *tangled*, *entangle* and *disentangle* all sound extremely similar because of a shared root. Words that sound similar may also tend to have similar meanings (Dautriche et al. 2017b; Monaghan et al. 2014) or syntactic functions (Kelly 1992), although form-meaning correspondences are generally very subtle; phonaesthemes are a notable exception (Bergen 2004). Historical relatedness is also a factor, since many words that map to distinct categories in their modern form trace their origins back to a shared ancestor; for example, *skirt* and *shirt* sound similar because they both come from the Old Norse *skyrta*. And finally, some adaptations to articulatory naturalness acting on individual words may also give rise to greater sound similarity *between* words; for example, words in languages with vowel harmony (Finley & Badecker 2008, 2010; A. Martin & Peperkamp 2020; A. Martin & White 2021) are more likely to share multiple segments.

Naturally, phonotactic constraints are also a major source of phonetic clustering: sounds and sound sequences that can appear in more contexts will be more frequent across a language[2]. Nonetheless, corpus analysis reveals a cross-linguistic tendency for lexicons to be *even* more clustered than required by the phonotactics of the language (Dautriche et al. 2017a). In particular, across a range of word lengths, high-frequency

[2]It is also worth noting that, while many phonotactic constraints appear to be phonetically-grounded and explainable in terms of substantive bias (Archangeli & Pulleybank 1994; Blevins 2004; Donegan & Stampe 2009; Finley & Badecker 2007; Hayes & White 2013; Moreton & Pater 2012; Zheng & Do 2025), it is also possible that some aspects of phonotactics arise from arbitrary responses to a pressure for clustering.

words tend to be more tightly clustered – both in terms of neighbourhood density and phonotactic probability – while lower frequency words tend to be more distinctive (Frauenfelder et al. 1993; King & Wedel 2020; Landauer & Streeter 1973; Mahowald et al. 2018; Meylan & Griffiths 2024). This pattern is suggestive of adaptation for efficient communication (Gibson et al. 2019; Jaeger & Tily 2011), since it minimises production effort for items that are produced most often, and maximises understandability for low-frequency items, which are often harder to process in comprehension (Brysbaert et al. 2018). More generally, the fact that lexicons are observably less disperse than they could be suggests that, overall, the advantages associated with word similarity outweigh the disadvantages. However, corpus data alone cannot provide causal evidence of a relationship between particular functional pressures and the structure of language.

In this study, we investigate how production and comprehension pressures compete to shape the degree of phonetic clustering in the lexicon. First, we set out an agent-based computational model of sound change (Section 4.3). In line with the psycholinguistic evidence reviewed above, we model production and comprehension pressures that pull in opposite directions. We test the prediction that natural-language-like lexicons will emerge only under the combined influence of both. In particular, we test whether clustered lexicons emerge, and whether this clustering is found particularly for high frequency words. To further explore the role of production and comprehension in shaping the lexicon, we then model a similar process in a behavioral experiment in which human participants communicate with a partner using a miniature artificial language (Section 4.4). To preview our results, the lexicons that emerged from our model when both production and comprehension pressures were at play were more clustered than those generated by comprehension pressures alone, but more disperse than those generated by production pressures alone. Similarly, in the experiment, manipulating the difficulty of only the production task or only the comprehension task gave rise to behaviours at one extreme or the other. When both tasks were difficult, participants adopted a variety of strategies, but overall there was more of a balance between ease of production and ease of perception. However, the effect of frequency on emergent lexicons was less clear; there was a subtle tendency in the model for more frequent words to become more clustered, but this pattern did not robustly materialise

in the experiment.

4.3 Computational model

We use an agent-based exemplar model (Nosofsky 1986; Wedel 2006) to test how mechanisms operating during individual episodes of production and comprehension might influence the degree of phonetic clustering present in a lexicon over time. In this model, pairs of agents use a miniature artificial language to communicate with each other over repeated rounds. In each communication round, agents take turns producing and interpreting signals, with some mechanisms that would be expected to favour or disfavour word similarity encoded within these processes (described in Section 4.3.1.3). Signals that result in successful communication are strengthened over time, while unsuccessful signals are more likely to drop out of the agents' memory. At the end of every round, we observe the state of the lexicon. The following section describes all of these components in detail; an overview is given in Figure 4.6. Readers wishing to skip the technical details can move on to Section 4.3.3 to see the results.

4.3.1 Details of the model

The model is implemented in Python 3.11; full code is available at https://osf.io/vsy6z/.

4.3.1.1 The agents

Each agent maintains their own independent internal representation of the lexicon, based on prior evidence. An agent's internal representation consists of 20 atomic meaning categories (represented by integers), each associated with a collection of signals. In the most basic version of the model, all meanings are equally frequent; we implement a simple frequency manipulation in Section 4.3.3.1. Each meaning category has a memory limit S (default value = 10) which constrains the number of signals that can be associated with it at any given time-point. When a new signal needs to be added to a category that is already at this limit, a random older signal is deleted first.

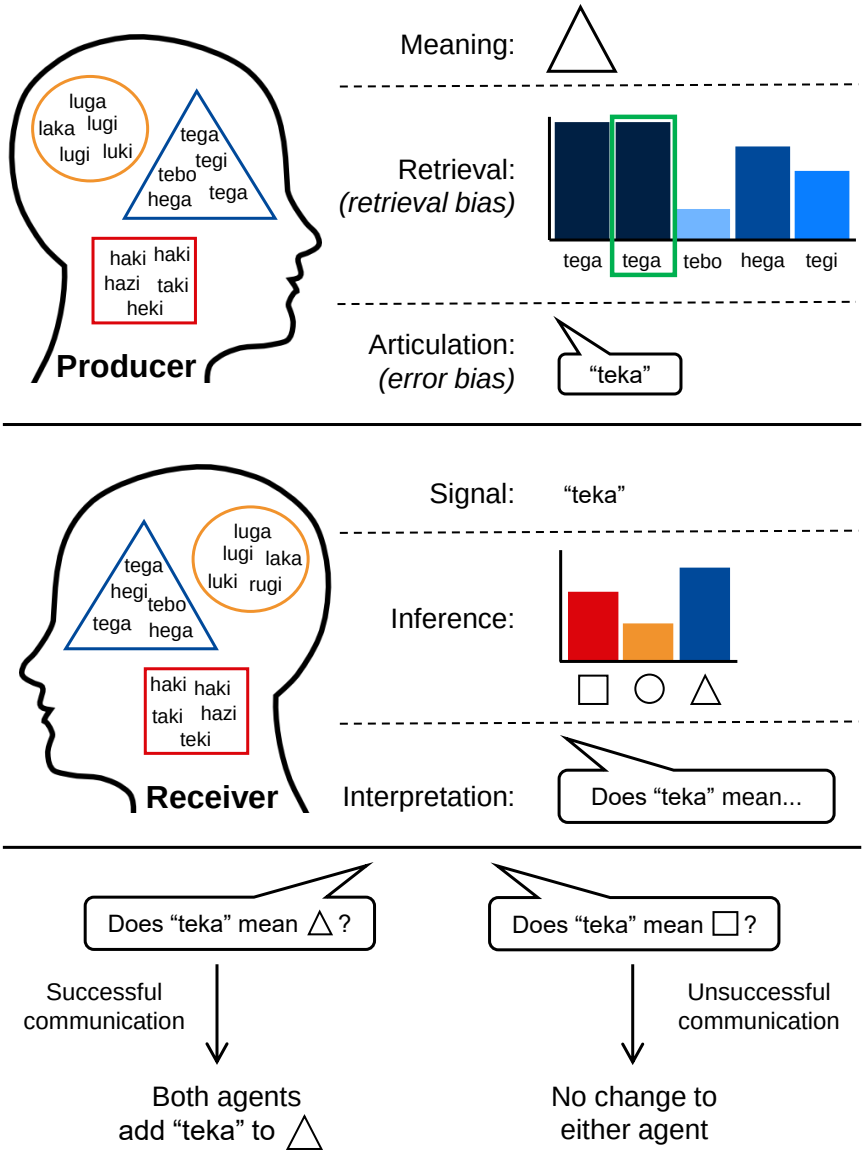

Figure 4.6: Overview of the model architecture for a single communication episode. Both agents maintain an independent internal representation of the lexicon in the form of meaning categories (shapes) and associated signals (exemplars). The Producer sends a signal to their partner to communicate about a target meaning, with two sources of similarity bias in this process. First, exemplars within the target meaning category are activated to different degrees depending on their phonotactic probability, meaning that exemplars that are more similar to others in the lexicon are more likely to be retrieved. Second, once an exemplar has been retrieved, there is some probability of an error being introduced into it during production; when an error is made, segments that are less frequent across the lexicon tend to be replaced by those that are more frequent. The Receiver compares the received signal to their stored exemplars to calculate a probability distribution over possible meanings, from which they sample a response; more distinctive signals give higher weight on the target meaning category relative to all other categories and are therefore more likely to result in successful communication, while signals that are more ambiguous between categories give a more uniform distribution over meanings and are therefore more likely to be misinterpreted. If the Receiver correctly infers the Producer's target meaning, both agents store the signal that was just sent as a new exemplar in that meaning category.

Since the model is exemplar-based, there is no abstract representation for agents to infer from the evidence they receive; rather, they store concrete exemplars of linguistic behaviour they've observed. As in Wedel (2012), we do not intend to make any claims about the specific nature of humans' mental lexicons[3]; this architecture is simply a convenient and transparent way to capture the fact that there is always fine-grained phonetic variation below the level of "the lexicon", and to show how this variation can provide the fodder for lexical evolution (Winter 2014). More specifically, while we might perceive words as having categorical boundaries, in reality, subtle variations in pronunciation mean that word boundaries are at least somewhat fuzzy, even within the same individual; different exemplars in our model can be thought of as representing this fuzziness.

4.3.1.2 The lexicon

The "words" agents store in our model are character strings Because we are interested in how clustering might emerge above and beyond the effects of word length (since shorter words are, necessarily, more similar to each other than longer words), word length is a constant in our model: all words are of length 8. For simplicity, the individual segments that make up a word are represented simply by letters, rather than by bundles of features or some other more phoneme-like representation (cf. Wedel 2012). Because of this simplification, it is not the case that segments can be more or less similar to each other: two segments are either identical, or they are different. Although this makes comparisons between words less nuanced, it is a reasonable simplification to improve model tractability, particularly given the lack of evidence that natural language lexicons are more clustered around highly distinctive contrasts than around more confusable contrasts (Dautriche et al. 2017a).

At the start of each run of the model, we generate 20 words (one per meaning category) by randomly combining letters from the set of English consonants. Letters are drawn from a uniform distribution, meaning that there is no pressure towards clustering coming from the initial lexicons. We use these words to seed a process of

[3] An alternative but related model could have been implemented in a Bayesian framework, with a compression-based prior (Kirby et al. 2015) that would favour lexicons with fewer unique sounds and sound combinations. However, such a model would locate clustering pressures in learning, whereas our primary interest here is the role of communication and use.

exemplar creation: specifically, the starting set of exemplars in each meaning category is a collection of S strings (where S is the memory limit for that category), each of which is created by randomly substituting a single character from the seed word assigned to that category. For example, if the seed word for a category was "tam", it could generate exemplars like "zam", "tum", and "tak".

Although agents therefore store a considerable amount of variation in their internal representation, we are treating exemplars as pronunciation variants of the same word, so we want to smooth out this within-category variation when we examine the state of the lexicon. To collapse an agents' internal representation down to a single word per meaning category — the canonical or 'average' form of the word — we simply concatenate the most common character in each position across all exemplars in that category. For example, given a set of exemplars {"miq", "mas", "taq", "maq"}, this process of concatenation would yield the word "maq", since "m" is the most common first letter, "a" is the most common second letter, and "q" is the most common final letter.

In order to analyse how the lexicon changes over time, and whether words are becoming more or less similar to each other, we calculate the *average pairwise edit distance* between words at each time step, including for the initial lexicon. Average pairwise edit distance, $D(L)$, is given by:

$$D(L) = \frac{\sum_{i,j \in L, i \neq j} LD(i,j)}{|L| \cdot (|L| - 1)} \tag{4.1}$$

where L is the lexicon, $|...|$ indicates cardinality (i.e. the number of words in L), i and j are words and $LD(i,j)$ is the Levenshtein distance between two words. That is, we calculate the edit distance between every pair of words in the lexicon, and then take the mean of these distances.

Because we generate the seed words randomly — so that all characters are equally likely to appear in all positions — words in the initial lexicon are always very different from each other: across 1,000 randomly generated lexicons, average pairwise edit distance had a mean value of 7.54 ($SD = 0.05$). In other words, in the initial lexicon, any two randomly selected words will usually differ at every position. If words are

becoming more similar to each other over time, this would be reflected by a *decrease* in average pairwise edit distance.

4.3.1.3 Communication

In each communication round, agents take turns as Producer and Receiver for all meanings. The Producer's task is to transmit a signal given a target meaning; the Receiver's task is to decode the intended meaning given a received signal. Whenever the Receiver successfully recovers the meaning of a signal, both agents store that signal as a new exemplar in the relevant meaning category. Due to the memory limit described in Section 4.3.1.1, exemplars that are either not used or do not result in successful communication will tend to drop out of the agents' internal representations over time.

Production Production consists of two stages: retrieval and articulation. In both of these stages, we build in observations from the psycholinguistic literature about how word similarity benefits word production. To summarise, exemplars that are more similar to others in the agent's internal representation are retrieved more easily (Chen & Mirman 2012; Goldrick & Larson 2008; Vitevitch 2002; Vitevitch et al. 2004), and errors in the pronunciation of a target exemplar tend to replace lower frequency segments with higher frequency ones (Dell 1986; Goldrick & Rapp 2007; Levitt & Healy 1985; Motley & Baars 1975; Munson 2001), thus creating sequences with higher phonotactic probability.

More specifically, production begins with the random choice of an exemplar from the target meaning category, where the probability of a particular choice depends on its phonotactic probability (average bigram positional probabilities across the string); exemplars with higher phonotactic probability are more strongly activated (the *retrieval bias* parameter). Before the exemplar is transmitted to the Receiver, an error is introduced into it with probability E[4]. All errors involve the substitution of a single segment

[4]In the simulations presented below, we use an unrealistically high E of 0.5, which would imply that language users mispronounce words around half the time. Using a larger E does not qualitatively change the results compared to a smaller E, but does allow effects to be seen in fewer time steps, which improves runtime. In any case, the function of the error mechanism is to introduce variation that can provide the fodder for lexical evolution; similar mechanisms in related models often apply to *every* production (e.g. Flego 2022; Wedel 2012; Wedel and Fatkullin 2017).

in a randomly chosen position. The new segment is sampled from the set of segments in the language, where the probability of selecting a particular segment depends on the frequency with which it occurs in the same context as the original segment across all exemplars in the agent's internal representation (the *error bias* parameter). By default, we only consider a single preceding segment when calculating conditional segment frequencies; in this way, errors tend to create high-probability bigrams. We use Laplace smoothing with parameter 0.01 to assign non-zero probability to segments that were present in the initial lexicon but have dropped out entirely, or segments that don't appear in a particular bigram. We also allow "substitution" to replace a segment with itself, which can happen when the segment targeted for error is very high-frequency in the given position; in this way, exemplars with high phonotactic probability in the language become less likely to be mispronounced.

Reception The final signal created by the Producer, including any error, is transmitted to the Receiver along with a context (list of possible meanings) which they have to choose from. The nature of this context is controlled by a *context size* parameter, which can take one of three values: maximal (the default: all meanings in the lexicon), random (n randomly selected meanings, where $1 \leq n \leq 20$), or minimal (= 1)[5].

When the Receiver hears a signal, they must infer its meaning by comparing it to all their stored exemplars for each meaning category in the current context. If the context contains only one meaning, the Receiver automatically assigns the signal to that meaning category. Otherwise, the probability of recovering the intended meaning is calculated using the Generalized Context Model (Nosofsky 1986, 2011)[6], which states that the probability of classifying stimulus i into category c_n is given by:

$$P(c_n|i) = \frac{\left[\sum_{j \in c_n} N_j \cdot \eta_{ij}\right]^\gamma}{\sum_{c \in C} \left[\sum_{k \in c} N_k \cdot \eta_{ik}\right]^\gamma} \tag{4.2}$$

where η_{ij} denotes the similarity between exemplars i and j and N_j is the frequency

[5]Using the minimal context size removes comprehension pressures from the equation entirely, since the Receiver has access to full information about the Producer's intended meaning, rendering their task trivial. A real-life analogue would be an utterance that takes place in a situation where there is only one salient possible interpretation. In our case, where communication is essentially just a process of object labelling, it could also be thought of as a Producer pointing at their intended referent.

[6]We exclude the category bias term used in the Generalized Context Model, since we want all categories to be equally likely *a priori*.

of exemplar j. The numerator is therefore simply the summed similarity score for the meaning category under consideration, and the denominator is the sum of all similarity scores for all meaning categories. γ is a response-scaling parameter which controls the Receiver's sampling behaviour: when $\gamma = 1$, the Receiver responds by sampling directly from the distribution of relative summed similarities over all categories (i.e. probability matching), whereas for higher values of γ, the Receiver responds more deterministically with the category that yields the largest summed similarity. Similarity between exemplars i and j is itself operationalised as the complement of the Levenshtein distance LD between the two strings, normalised by dividing by M, the length of the longer string[7]:

$$\eta_{ij} = 1 - \frac{LD(i,j)}{M} \tag{4.3}$$

The Receiver samples a meaning from the context using the relative similarity scores given by Equation 4.2 as weights. The effect of this reception mechanism is that more distinctive signals will be more likely to result in successful communication, since they will give higher weight on the target meaning category relative to all other categories. On the other hand, signals that are similar to exemplars in multiple categories will give a more uniform distribution over possible meanings, and are therefore more likely to be misinterpreted.

4.3.1.4 Iteration

At the end of every communication round, we extract the current state of the lexicon from one of the agents (randomly chosen) and calculate its average pairwise edit distance, $D(L)$. A new communication round then starts; each run of the model consists of 4,000 such rounds. Note that there is no transmission of the language to naive individuals between communication rounds (cf. Kirby et al. 2015); the same pair of agents continue to communicate with each other throughout the simulation. Since there are no learning biases in this model, the only purpose of including naive agents would be to introduce a source of random drift, which is already provided by limiting our agents' memory capacity (Spike et al. 2013, 2017).

[7] M is a constant in this case, since all words in our model are the same length.

4.3.2 Simulations

We use the model to run simulations in three conditions:

- **Production pressures only:** Both the *retrieval bias* and *error bias* parameters are switched on, but *context size* is set to minimal, such that there is no inference on the Receiver's part and communication is always successful.
- **Comprehension pressures only:** *Context size* is set to maximal, requiring the Receiver to compare received signals to exemplars in all possible meaning categories to determine the Producer's intended meaning. However, both the *retrieval bias* and *error bias* parameters are switched off: all exemplars have equal probability of being retrieved for production, and errors simply replace one random segment with another random segment.
- **Competing pressures:** Both the *retrieval bias* and *error bias* parameters are switched on, and *context size* is set to maximal.

For the latter two conditions, we also test a range of different values for the Receiver's γ parameter (which influences how deterministically they choose the meaning category that best fits the received signal). For each configuration of parameter settings, we run 10 simulations — each with a different random input lexicon and set of starting exemplars.

4.3.3 Results

Recall that the measure of similarity we use here is *average pairwise edit distance*, $D(L)$. When average pairwise edit distance is lower, it mean that words are more similar to each other. Figure 4.7 shows the change in average pairwise edit distance over time in three conditions. When only production pressures are present, the Producer's similarity biases completely take over: lexicons become rapidly more clustered, often to the point of *degeneracy* (Kirby et al. 2015), where there is just one word for every meaning ($D(L) = 0$). Conversely, when comprehensibility is the only pressure on the language, lexicons remain very disperse over time.

When there is competition between similarity biases in production and the pressure for distinctiveness arising from communication, the result is a more balanced lexicon: words are somewhat more clustered together, but not to such an extreme degree (i.e. degeneracy) as in the production-only condition. The speed with which clustering increases depends on the strength of the comprehension-side pressure for distinctiveness, controlled by the Receiver's γ parameter: when γ is higher, the pressure for distinctiveness is weaker, which allows lexicons to change more rapidly. However, the curve eventually flattens out; this plateau can be thought of as the state in which words are as similar to each other as they can be whilst still allowing the Receiver to tell them apart with a reasonable level of accuracy.

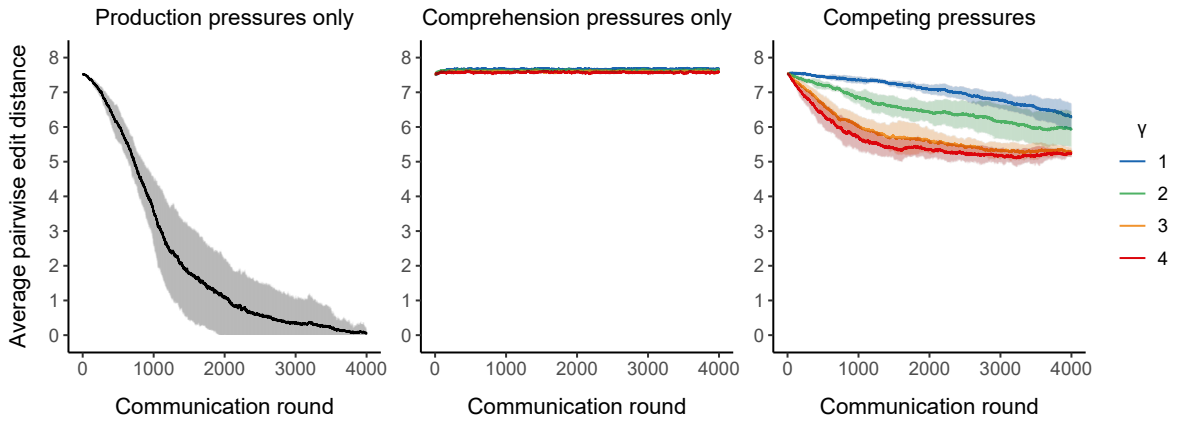

Figure 4.7: Average pairwise edit distance over 4,000 communication rounds in three conditions; lower numbers mean that words are more similar to each other. Bold lines represent the mean across 10 runs; shaded areas around these lines represent ±1 standard deviation. Colours in the two right-hand plots represent different values of the Receiver's γ parameter, which controls the strength of the comprehension-side pressure for distinctiveness; higher values correspond to a weaker distinctiveness pressure. With production pressures alone, lexicons rapidly degenerate. With comprehension pressures alone, lexicons remain in their starting state, where words are all very different from each other. Only with competition between production and comprehension pressures does an intermediate state emerge, in which lexicons become somewhat more clustered but ultimately stabilise.

Overall then, when we allow lexicons to be shaped by only one aspect of communication, the results are extreme and bear little resemblance to natural languages. Words either become so similar that they cannot be distinguished at all (production-only), or they remain totally dispersed (comprehension-only). It is only when both pressures are present — as they are in real communication — that a middle ground emerges.

4.3.3.1 Adding frequency effects

As described in Section 4.2, the degree of clustering is not the same across all parts of natural language lexicons: more frequent words tend to be more similar to each other, while lower frequency words tend to be more distinctive (Frauenfelder et al. 1993; King & Wedel 2020; Landauer & Streeter 1973; Mahowald et al. 2018; Meylan & Griffiths 2024). In the model results described above this effect is of course not observable, since all meanings were equally frequent. Next, we incorporate a simple notion of frequency to test whether the effect of frequency emerges from the model. Specifically, we assign 5 meanings to a high-frequency group, and the other 15 to a low-frequency group. During each round, agents communicate about the high-frequency meanings three times as often as the low-frequency meanings (three trials per agent per high-frequency meaning, versus one for the low-frequency meanings). Additionally, we increase agents' memory limit for high-frequency meanings to 30 (the memory limit for low-frequency meanings stays at 10) to capture the fact that high-frequency lexical items have stronger mental representations than their low-frequency counterparts (Alexandrov et al. 2011; Popov and Reder 2020; see also the multiple-trace hypothesis: Hintzman and Block 1971). The rest of the model architecture is identical.

Figure 4.8 shows the change in average pairwise edit distance over time in the same three conditions as above, now additionally split by frequency. The results for the first two configurations look very similar as in Figure 4.7, with no difference between frequent and infrequent words: lexicons remain in their starting state in the comprehension-only condition, and rapidly degenerate in the production-only condition. However, crucially, when production and comprehension pressures are in competition, there is a very subtle effect of frequency. Specifically, clustering increases slightly more on average in the high-frequency component of the lexicon, but only when the Receiver's γ parameter is low; this suggests that the benefits conferred by increased frequency (due to having a stronger mental representation for higher frequency items) are washed out when the Receiver is already very proficient at telling words apart.

The effect of frequency becomes more apparent if we make two further modifications to the model architecture. First, we can modulate the strength of the producer biases such that they are stronger for higher frequency words. For example, in the

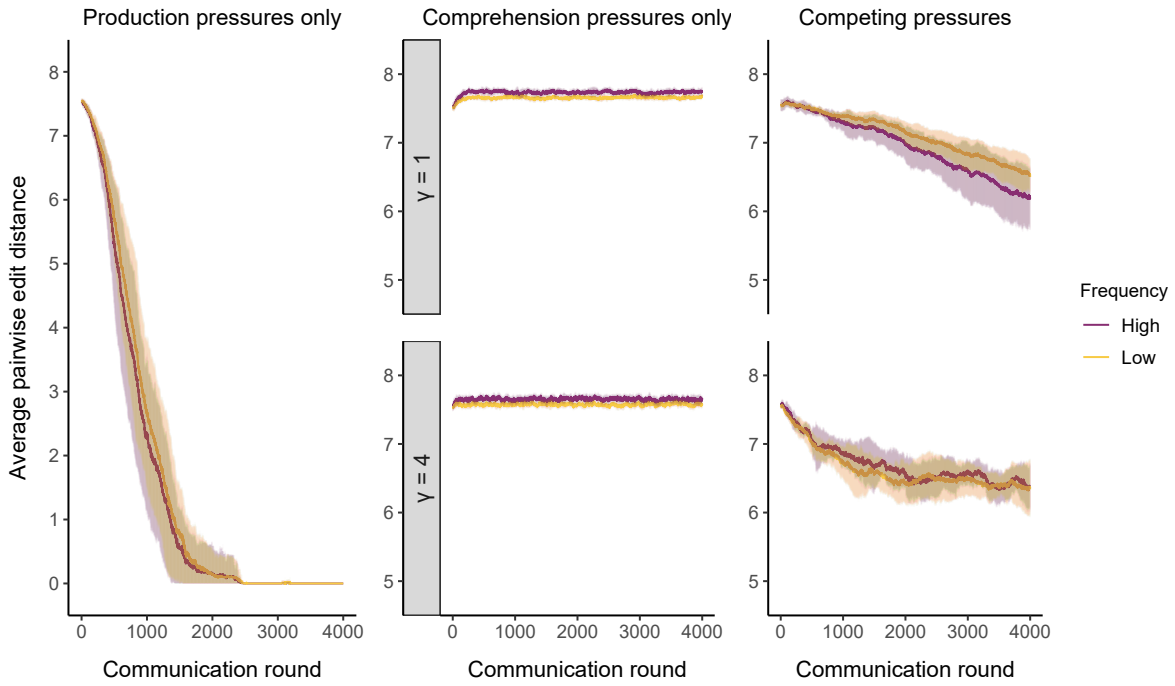

Figure 4.8: Average pairwise edit distance for the high and low-frequency components of the lexicon over 4,000 communication rounds. With only production pressures, lexicons rapidly degenerate, with no difference between frequent and infrequent words. With only comprehension pressures, both high and low-frequency words remain very distinct over time. When both production and comprehension pressures are present, a very subtle effect of frequency emerges: the high-frequency component of the lexicon becomes slightly more clustered than the low-frequency component, but only when the Receiver's γ parameter is low (top).

case of word length, there is good evidence that speakers preferentially shorten high-frequency words (e.g. Bybee 2002; Kanwal et al. 2017; Mahowald et al. 2013; Pierrehumbert 2001). We can encode a similar preference to maximise ease-of-production for high-frequency items in our model by raising the activation values given by the Producer's *retrieval bias* parameter (described in Section 4.3.1.3) to the power of 2 when they are labelling a high-frequency meaning. This has the effect of exaggerating the preference for exemplars with high phonotactic probability. Second, we can treat high-frequency words as requiring less inference by the Receiver. The logic here is that high-frequency meanings will be weighted more highly *a priori*, so if a received signal is a good fit to a high-frequency category, the Receiver might not consider as many alternatives (note also that high-frequency words attract more attention early in processing: Dahan et al. 2001). We can operationalise this intuition by manipulating the *context size* parameter (described in Section 4.3.1.3): for high-frequency items, the Receiver only

has to choose between 5 candidate meanings, while for low-frequency items, there are 15 candidate meanings. Figure 4.9 shows the results of this model configuration when production and comprehension pressures are in competition[8]. Here, the effect of frequency is much clearer: the high-frequency component of the lexicon becomes more clustered more quickly than the low-frequency component. However, again, this effect is only observable for lower values of the Receiver's γ parameter.

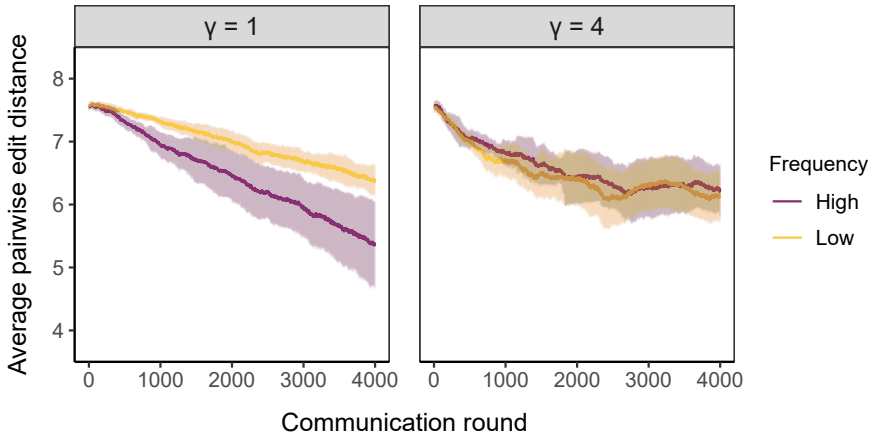

Figure 4.9: Average pairwise edit distance for the high and low-frequency components of the lexicon when production and comprehension pressures are in competition, with two additional modifications to the model architecture: (1) Producer biases are stronger for high-frequency items, and (2) high-frequency items are more predictable for the Receiver. In this configuration, an effect of frequency is evident when the Receiver's γ parameter is low (left), but still does not emerge for higher values of γ (right).

4.3.4 Model discussion

Our model shows that phonetic clustering — a robust property of natural language lexicons — can emerge from initially random languages during repeated episodes of communication. Specifically, moderately-clustered lexicons emerge when there is competition between production pressures (which favour greater similarity between words) on the one hand, and comprehension pressures (which favour greater distinctiveness) on the other. With just one or other of these pressures, lexicons tend to fall within an extreme region of the possible design space: under the influence of production pressures alone, lexicons degenerate to the point of being communicatively useless, while when comprehension is the only pressure, lexicons remain in their initial, maximally disperse state.

[8]We only show this condition here since we have already established that there is no effect of frequency in the other two conditions.

Although models are always a simplification of the system they are designed to study, it is worth revisiting the specific simplifying assumptions we have made here. Firstly, as described in Section 4.3.1.2, we do not use a feature-based representation of the segments within a word, unlike in some similar models (e.g. Wedel 2012). Such a model architecture would probably improve the Receiver's performance, by allowing them to make more sophisticated comparisons between a received signal and their stored exemplars. However, since such fine-grained patterns of similarity do not feature in the calculations of phonotactic probability and bigram frequency that drive the Producer's behaviour, we do not think there would be significant downstream consequences for the eventual outcome of the model. Rather, clustering would likely just emerge *faster* since greater success on the Receiver's part results in more frequent storage of new exemplars and quicker turnover of old exemplars. In any case, corpus analysis suggests that a feature-based representation is unnecessary to explain the degree of clustering in natural language lexicons (Dautriche et al. 2017a), which is the basis on which we made this simplification.

Furthermore, whilst successful communication changes the agents' internal representation, there is no such feedback loop from unsuccessful communication in the model. This is a common feature of exemplar models in this tradition (e.g. Wedel 2012; Wedel & Fatkullin 2017), since there is no penalty on unsuccessful signals (beyond not being stored in the target category) encoded within the Generalised Context Model of signal reception (Nosofsky 1986, 2011). However, other frameworks exist that could capture the intuition that language users might try not to use variants that they do not believe to be communicatively useful. For example, various types of models employ some kind of negative feedback after unsuccessful interactions, either deletion or inhibition as in reinforcement models (e.g. Barrett 2006; Franke & Jäger 2012; Skyrms 2010) or weakening associations as in the Naming Game (Steels 2012; Steels & Loetzsch 2012); for further discussion of these mechanisms, see Spike et al. 2017. However, the decision about how to implement such mechanisms is not straightforward, especially in the case of signals containing errors whereby there is no exactly matching exemplar in either agents' internal representation that could be targeted. An alternative to penalising signals after communication has failed is to downweight signals that are more likely to result in failure *before* an interaction takes place, as in the Rational

Speech Act (Frank & Goodman 2014; Goodman & Frank 2016); in such models, a pragmatic speaker reasons about how likely a listener would be to recover the intended meaning from the different utterances available to them. The downside of this kind of mechanism is that it requires a significant amount of computation in every communication episode, dramatically increasing the runtime of the models. Listener-oriented approaches have also been criticised as teleological (e.g. Wedel 2006). In any case, we would argue that either of these approaches adds unnecessary complication to the model; selection of successful signals works by itself, it simply takes slightly longer to turn over less useful signals.

Finally, it is true that comprehension does not straightforwardly favour word dissimilarity, as suggested by our model of reception: specifically, increases in phonotactic probability have been found to facilitate word recognition (Vitevitch & Luce 1998). However, pure recognition — in terms of deciding whether a received stimulus is familiar (word) or unfamiliar (non-word) — is very different from the categorisation task faced by our agents, a task where competition between multiple activated referents is known to inhibit processing (Luce & Pisoni 1998). Indeed, Vitevitch and Luce (1998) describe the effect of phonotactic probability as facilitative for sub-lexical processing (for example, segmenting the speech stream, or processing novel sound sequences) and inhibitory for lexical processing (for example, determining the intended meaning of a received signal, as in our model). Wedel (2012) also points out that the general behaviour of these exemplars models is the same whether similarity biases are encoded once (in production) or twice (in production and perception).

Returning to the frequency effects discussed in Section 4.3.3.1, our results suggest that frequency may modulate the rate of lexical evolution, with the effect depending to some extent on the assumptions we make about the processing consequences of frequency. In the most basic version of our frequency manipulation, we implicitly assume that production biases are underlyingly frequency-*in*dependent. In other words, the model architecture is such that producers want to maximise production ease across the board; frequency-dependent lexical evolution emerges simply because they can get away with doing so more for high-frequency items. The fact that frequency effects are so subtle under this assumption makes sense when we examine how frequency actually impacts the two participants in a conversation. From the comprehender's

side, a frequency advantage is baked into the reception mechanism (Equation 4.2): the stronger mental representation of high-frequency items (due to their larger memory limit) increases the Receiver's certainty that a received signal maps onto a target category. However, from the producer's side, any selection which may be acting to change a word's form is competing against the fact that the representation of the word's existing form is very strong; this may also be why, for example, high-frequency irregular items tend to resist regularisation (e.g. Bybee 1995; Cuskley et al. 2014; Sims-Williams 2022; K. Smith et al. 2023; Wu et al. 2019). Therefore, while comprehension may permit greater clustering for high-frequency items, the production process may be slower to generate the variation required for selection to act upon for these items. A stronger effect of frequency can emerge from the model under certain conditions, but of course, it may not be desirable to make the additional assumptions required to generate this result (Marquet et al. 2014). Future work could expand upon the frequency aspect of our model, for example, by using a more realistic distribution of word frequencies (i.e. following a power law) rather than treating frequency as a binary value.

Overall though, our model predicts that production or comprehension pressures in isolation will give rise to lexicons at one extreme of clustering or the other. An intermediate state, with levels of clustering more similar to those found in natural language lexicons, should emerge when these pressures are in competition. In the next section, we simulate these same pressures in a communication experiment with human participants, focusing more specifically on the interaction between clustering and frequency.

4.4 Communication experiment

We use an artificial language learning paradigm to investigate how production and comprehension pressures trade-off against each other to influence language users' lexical choices during communication. The experiment is inspired by Kanwal et al. (2017), who showed that the Law of Abbreviation (Zipf 1949) emerges from precisely such a trade-off. Specifically, in their experiment, participants were trained on a miniature lexicon in which two objects that differed in frequency were labelled with either a unique,

long label ("zopudon" or "zopekil") or a shared (and therefore ambiguous) short label, "zop". Kanwal et al. found that participants favoured the ambiguous short label (which was quicker to produce) under time pressure, and the unambiguous long labels under pressure for accuracy. When both of these pressures were present, participants converged on an optimal solution, whereby the short label was consistently mapped to the high-frequency object and the long label to the low-frequency object, consistent with the Law of Abbreviation. By simulating the pressures inherent to real communication, this method provides a convenient way to disentangle the individual effects of opposing pressures, and to show that key structural properties of natural languages can emerge from their confluence.

Following Kanwal et al., rather than relying on participants to introduce changes to the lexicon themselves — i.e. make errors in production — we designed a lexicon incorporating lexical variation. However, the competitors in our experiment are words from different phonological neighbourhoods, rather than words of different lengths. Specifically, each object was labelled by two different words: one from a high-density neighbourhood (highly confusible with words belonging to other meanings), and one from a low-density neighbourhood (highly dissimilar from all other words in the language). As in Kanwal et al., participants were trained on the different names for two objects that differed in frequency, and were then paired up to play a communication game, during which we manipulated the presence or absence of a production-side pressure for similarity (Stemberger 2004; Vitevitch & Luce 2005; Vitevitch & Sommers 2003) and a comprehension-side pressure for distinctiveness (Chan & Vitevitch 2009; Luce & Pisoni 1998). We predicted that natural-language-like properties would arise only when both these pressures were present.

4.4.1 Methods

The study was approved by the PPLS Ethics Committee at the University of Edinburgh (ref. 6-2425/1) and was pre-registered with the Open Science Foundation (https://osf.io/jucn6).

4.4.1.1 Materials

The meaning space consisted of two objects — a compass and a lightbulb — represented by drawings from the MultiPic databank (Duñabeitia et al. 2018). The two drawings score very similarly for visual complexity (2.65 and 2.41 respectively, on a scale from 1 to 5). To investigate the role of frequency on clustering, one object (randomly chosen for each participant) appeared three times more frequently than the other throughout the experiment. The language consisted of four artificial CVC words: "zun" /zʌn/ and "zan" /zæn/ (the *high neighbourhood density* words; henceforth, HND) and "mig" /mɪg/ and "tep" /tɛp/ (the *low neighbourhood density* words; henceforth, LND). The artificial words are matched for neighbourhood density in English (56 ± 1) according to the CELEX corpus (Baayen et al. 1995) and have average positional phoneme probability ranging between 0.0498 and 0.0583 according to the Irvine Phonotactic Online Dictionary (Vaden et al. 2009). We designed the words in this way to ensure that any preference for either HND or LND words would be driven only by their status within the artificial language, not by their relationship to participants' native English. Audio files for each word were synthesised using an online IPA to Speech tool (https://www.antvaset.com/ipa-to-speech). For each participant, each object was randomly assigned two names: one from each neighbourhood. Unlike in Kanwal et al. (2017), the competitor labels for an object were therefore not variants of a single word (e.g. "zopudon" → "zop"), but two completely different words. We designed the lexicon in this way to maximise the distance between the LND words: any words that were more clearly derived from the HND words would necessarily also be quite similar to each other, reducing their distinctiveness.

4.4.1.2 Procedure

The experiment was written in JavaScript using the jsPsych library (Leeuw et al. 2023). The design is based on the paradigm developed by Kanwal et al. (2017). A schematic of the experimental design and procedure is given in Figure 4.10. Participants completed the following phases, in the order shown below.

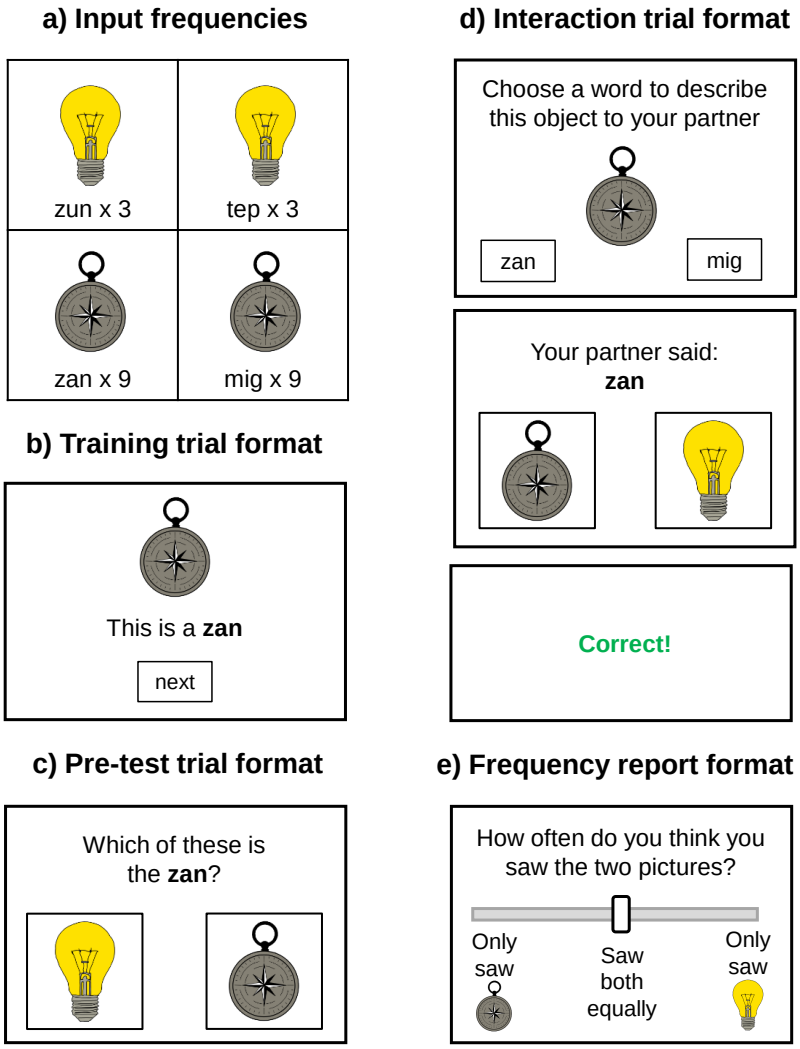

Figure 4.10: Schematic of the experimental design and procedure. (a) Example training set (the exact permutation of objects and labels was randomised for each participant) showing the 75/25 frequency distribution over the two objects (rows) and 50/50 distribution over HND and LND words (columns). (b) Example training trial. (c) Example pre-test trial. (d) Example interaction trial, proceeding from a Director trial (top) to a Matcher trial (middle) and then feedback to both participants (bottom). (e) Example frequency report trial.

Training On each training trial, an object was presented on screen alone for 1000ms while the audio file of the appropriate word played once. The orthographic form of the word then appeared below the image in the English frame 'This is a . . .'. After another 1500ms, a 'next' button appeared to let participants advance to the next trial. Participants completed 24 training trials: 18 for the frequent object, and 6 for the infrequent object. Each object appeared half the time with its HND word and half the time with its LND word. The order of training trials was randomised for each participant.

Pre-test After the training phase, participants were tested on their knowledge of the language. On each trial, participants were presented with a word from the artificial language in the English frame 'Which of these is the . . .?' and asked to choose between the two objects. They received full feedback on their response. Again, participants completed 24 trials, with the same distribution over frequent/infrequent meanings and HND/LND words as in training. The order of trials was randomised for each participant. Participants were required to reach at least 83% accuracy (i.e. ≥ 20 trials correct) to proceed to the interaction phase. Additionally, two attention checks were randomly interspersed within this phase. On these trials, participants saw a familiar English word in the same 'Which of these is the . . .?' frame, along with two previously unseen pictures. They received no feedback on their response to these trials. Participants were required to pass at least one of these attention checks to proceed to the interaction phase.

Interaction The interaction phase of the experiment was managed via a Python WebSockets server (based on code from https://kennysmithed.github.io/oels2023/[9]). At the start of the interaction phase, participants were put into a virtual waiting room ready to be paired with the next participant who completed the pre-test. An on-screen timer kept participants informed of how long they had been waiting. If participants were not paired with a partner within 5 minutes, they were removed from the waiting room and paid for their time.

Once participants were paired, they played a communication game. Participants were instructed that they had two goals: to score as many points as possible (i.e. the *accuracy* pressure in Kanwal et al. 2017) and to complete the game as quickly as possible (i.e. the *time* pressure in Kanwal et al. 2017).

On each trial, one participant acted as the Director and the other as the Matcher; roles alternated between every trial. The Director was shown an object and asked to name it for their partner. An on-screen stopwatch tracked how long the Director took to complete this task (to reinforce the pressure for speed). The Director was always given both object names as options, but the method of producing a word differed between conditions, as outlined below. The Matcher was shown the word sent by the Director

[9]Full code for the experiment is available at https://osf.io/vsy6z/.

(with or without noise depending on condition; see below) and asked to choose which object they thought their partner was describing. Both participants received feedback as to whether the Matcher chose the correct object (to reiterate the pressure for accuracy). Participants completed 24 trials as Director and 24 as Matcher, with the same distribution over frequent/infrequent meanings as in training. The order of each participant's Director trials was randomised. At the end of the interaction phase, both participants were shown their pair's final score and overall completion time.

To avoid having to ensure that participants were trained on the same version of the input language (since the assignment of objects to frequencies and words to objects was randomised for each participant), participants' responses were translated via a shared underlying representation before being transmitted, following a similar method to that used by K. Smith et al. (2024). Specifically, if the object being labelled by the Director was the high-frequency object in their training set, then the target object (i.e. correct answer) for the Matcher would be whichever object was the high-frequency object in *their* training set. Similarly, if the Director sent the HND word that they were trained on for their target object, then the Matcher would see the HND word that *they* were trained on for *their* target object (i.e. the object of the same frequency as the object seen by the Director). This procedure is illustrated in Figure 4.11.

Each pair was randomly assigned to one of the three experimental conditions. There were two different versions of the Director and Matcher trials — an easy version, and a more difficult version — depending on condition. In the PRODUCTION condition, Director trials were difficult but Matcher trials were easy. In the COMPREHENSION condition, it was the other way around: Matcher trials were difficult but Director trials were easy. In the critical COMBINED condition, both tasks were difficult. Specifically, the manipulations were as follows (also illustrated in Figure 4.12):

- **Easy Director trials:** The Director was presented with both word options for the target object (in a random order) and simply asked to click on the word they wished to send.
- **Difficult Director trials:** The Director was presented with both word options for the target object (in a random order) and asked to use a 3x6 on-screen keyboard to type one of the words. They were only able to transmit one of the valid words;

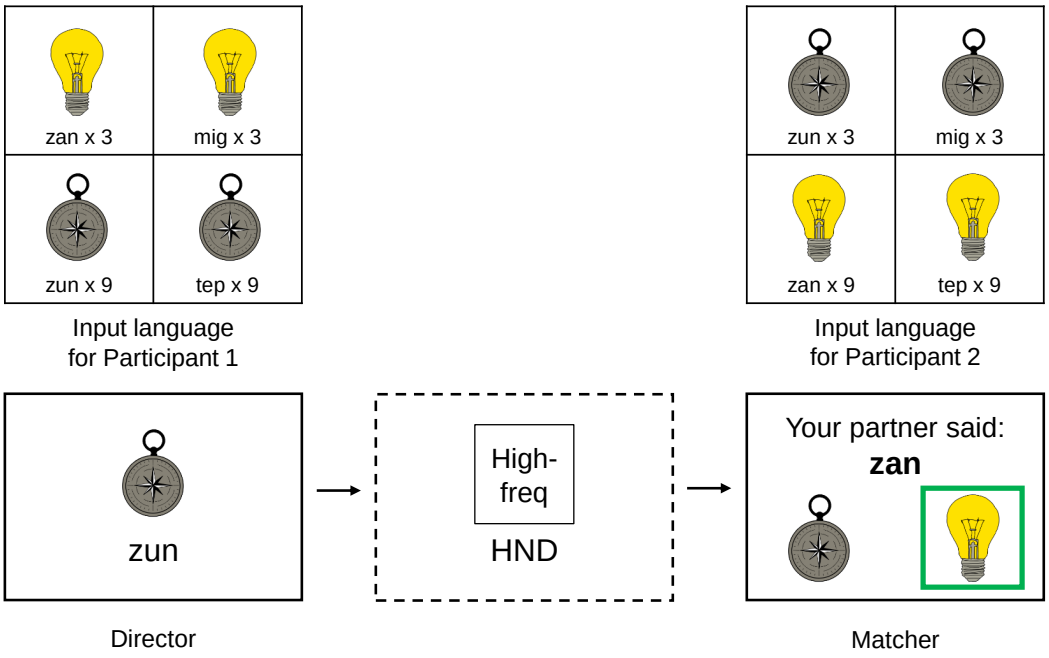

Figure 4.11: Example of the procedure for transmitting responses in the interaction phase between two participants who were trained on a different random permutation of the input language. The Director sees the compass (which was the high-frequency object in their training set) and sends the word "zun". This is first translated into an underlying representation whereby objects are represented by their frequency and words by their neighbourhood, rather than either being associated with specific forms. This underlying representation is then used to determine which word form to show the Matcher and which object should be the target; in this case, the lightbulb is the target object since this was the high-frequency object in the Matcher's training set, and its associated HND word is "zan".

if they submitted a word that didn't exist in the artificial language, or that referred to the other object, they were asked to try again[10]. The letters required to make an HND word ("z", "u", "a" and "n") always appeared in the same positions in the centre of the keyboard. The letters required to make an LND word ("t", "e", "p", "m", "i" and "g"), along with six other distractor letters that were not used in the artificial language, appeared around the outside of the keyboard and changed positions on every trial. Additionally, the central four buttons were three times as large (both in area and in font size) as the outer buttons. In this way, HND words were easier to produce than LND words. This design was intended to simulate the idea that, in spoken word production, frequently-used phonemes are pronounced more quickly and accurately, while less frequently-

[10]We included this restriction for two reasons. Firstly, the translation procedure illustrated in Figure 4.11 would only work if it was possible to definitively map participants' responses to categories from the input language. And secondly, the Matcher in the COMPREHENSION condition would always see a valid word since the Director had no freedom to invent new forms, so we wanted to ensure that this aspect was parallel across conditions.

used phonemes present more of a moving target for pronunciation (Goldrick & Larson 2008; Goldrick & Rapp 2007; Munson 2001; Vitevitch et al. 2004).

- **Easy Matcher trials:** Transmission was clean, and the Matcher was presented with the full word sent by the Director (after any necessary translation; see above).
- **Difficult Matcher trials:** Transmission was noisy, and the Matcher was presented with only the first letter of the word sent by the Director (after any necessary translation; see above). One letter provided enough information to distinguish between the LND words, but this information loss rendered the HND words identical and therefore ambiguous between the two objects. This design was intended to simulate the idea that, in spoken word perception, words with many neighbours activate many candidate meanings, and are thus more likely to be misinterpreted, while more distinctive words are more likely to activate only the target meaning (Chan & Vitevitch 2009; Luce & Pisoni 1998).

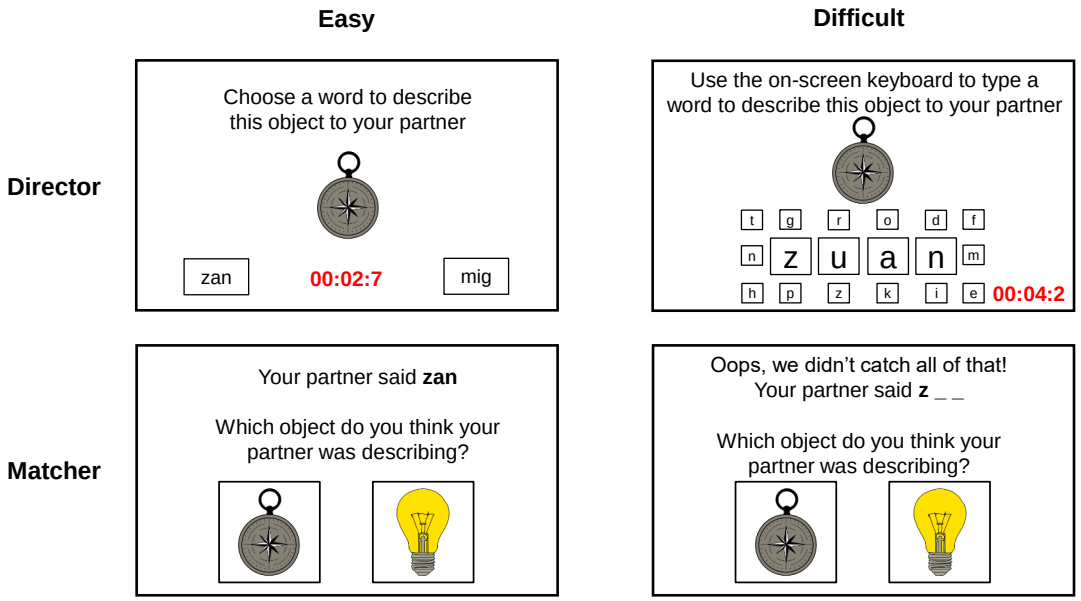

Figure 4.12: Easy (left) and more difficult (right) versions of the Director (top) and Matcher (bottom) tasks. When the tasks are easy, HND and LND words are similarly easy to produce and comprehend. When the tasks are difficult, there is a production-side pressure in favour of HND words, which are made up of more accessible segments, and a comprehension-side pressure in favour of LND words, which are able to overcome the noise on transmission.

Frequency report Once participants completed the interaction phase, they were asked to complete one final task individually. This task was included as a sense check that

participants had noticed the frequency imbalance between the two objects. Participants were presented with a continuous slider over percentages and asked "How often do you think you saw the two pictures? Did you see one more than the other?". The slider was accompanied by three labels: "Only saw *Object 1*" at one end, "Saw both objects equally often" in the middle, and "Only saw *Object 2*" at the other end. Which object appeared at which end of the slider was randomised for every participant.

4.4.1.3 Participants and exclusions

We used Prolific to recruit 220 adults resident in the UK who self-reported that their first language was English and that they had no known language disorders. They were provided with a downloadable information sheet and gave informed consent to participate. The experiment took around 20 minutes to complete in full (median time = 17:46), for which participants were paid £3.50 (above UK National Minimum Wage at the time of running the experiment). Seven participants were prevented from proceeding to the communication game due to low accuracy on the pre-test[11]; these participants were paid a reduced rate of £1.75. 27 participants started but failed to complete the interaction phase (either due to technical difficulties during the communication game or because they timed-out of the waiting room before being paired with a partner); these participants were paid a variable rate depending on how far they had got through the experiment. Six participants (one pair in each condition) completed the communication game and were paid the full rate, but their data was excluded from analysis because their completion time was more than 3 standard deviations above the median in that condition. We also pre-registered that we would exclude data from participants who admitted to taking written notes in a debrief questionnaire; no participants were excluded on this criterion. After all exclusions and dropouts, we were left with 30 pairs in each condition: a total of 180 individual participants.

[11] All participants passed both attention checks, so these exclusions were all due to low accuracy on critical trials.

4.4.1.4 Predictions

We predicted that participants in the PRODUCTION condition, where HND words were easier to produce than LND words, would tend to use the HND word for both objects, regardless of frequency. By contrast, we predicted that participants in the COMPREHENSION condition, where noisy transmission meant that HND words (but not LND words) became indistinguishable, would tend to use the LND word for both objects, regardless of frequency. We predicted that we would observe a natural-language-like frequency trade-off in the critical COMBINED condition, where both these pressures were present, such that participants would consistently map the frequent object to the HND word and the infrequent object to the LND word. This is the optimal strategy by which to minimise production effort (and therefore complete the game as quickly as possible) but still maintain an unambiguous one-to-one form-meaning mapping (and therefore score as many points as possible).

4.4.2 Results

4.4.2.1 Confirmatory analysis

Figure 4.13 shows the proportion of trials on which each pair used the HND word on Director trials, split by object frequency and condition. As predicted, most participants in the COMPREHENSION condition used the LND word for both objects, while in the PRODUCTION condition, most participants used the HND word for both objects. In the critical COMBINED condition, where the HND words were considerably easier to produce for the Director but functionally ambiguous for the Matcher, participants adopted a range of strategies. Some arrived at the optimal strategy described in Section 4.4.1.4. However, many were willing to expend extra time and effort to use the LND words for both objects and thus ensure accurate communication, while others opted to use the HND words for both objects and thus minimise transmission time at the expense of perfect accuracy.

We used the `lme4` package (D. Bates et al. 2015) in R (R Core Team 2024) to fit a logistic mixed effects model to the data, with a binary dependent variable of HND word use (as contrasted with LND word use, i.e. 1 if the participant produced the

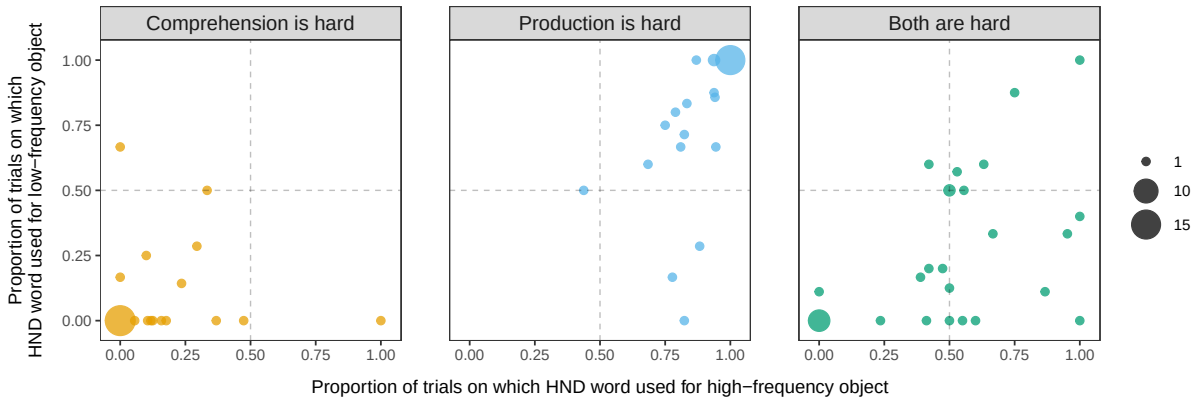

Figure 4.13: Proportion of trials on which the HND word was used for the high-frequency object vs. the proportion of trials on which it was used for the low-frequency object. Each data point combines a pair of communicating players, representing the sum of their Director trial productions. As in Kanwal et al. (2017), only data from the second half of each pair's interaction trials is shown, as participants were more likely to have converged on a stable mapping by this time. Data points in the bottom left quadrant indicate pairs who are mostly using the LND words for both objects; participants are clustered in this quadrant in the COMPREHENSION condition (left), where only the LND words are reliably distinguishable and there is no countervailing pressure from production in favour of the HND words. Data points in the top right quadrant indicate pairs who are mostly using the HND words for both objects; participants are clustered in this quadrant in the PRODUCTION condition (middle), where HND words are considerably easier to produce than LND words and there is no countervailing pressure from comprehension in favour of the LND words. Data points in the bottom right quadrant indicate pairs who are mostly using the HND word for the frequent object and the LND word for the infrequent object. This behaviour, consistent with the frequency trade-off seen in natural languages, is numerically most common in the critical COMBINED condition (right), where both production and comprehension pressures are at play, but a range of other behaviours are also represented in this condition.

HND word, 0 if they produced the LND word). The model included fixed effects of experimental condition (treatment-coded with the COMPREHENSION condition as the reference level), object frequency (treatment-coded with low-frequency as the reference level) and their interaction, and nested by-participant and by-pair random intercepts and random slopes for object frequency [12]. As in Kanwal et al. (2017), only data from the second half of each participant's Director trials was included in the model, as pairs were more likely to have converged on a stable mapping by this time. The model reveals that participants in the COMPREHENSION condition were very unlikely to use the HND words for either object, while participants in the PRODUCTION condition were very likely to use the HND words for both objects. The predicted interaction between condition and frequency was not statistically significant, meaning that there is insufficient evidence to conclude that participants in the critical COMBINED condition were

[12]Model formula: `HND word ~ condition + frequency + condition:frequency + (frequency | pair/participant)`

Table 4.1: Summary of fixed effects for a logistic mixed effects model with HND word use as the binary dependent variable, and nested by-participant and by-pair random effects for object frequency. The predicted effects are shown in bold. Coefficient estimates are on the log-odds scale.

	β	SE	z	p
intercept (object = infrequent, condition = Comprehension)	**-8.075**	**1.590**	**-5.078**	**<0.001**
object = frequent	0.807	1.707	0.473	0.636
condition = Production	**14.024**	**2.526**	**5.553**	**<0.001**
condition = Combined	3.893	1.434	2.714	<0.01
object = frequent & condition = Production	0.582	2.787	0.209	0.835
object = frequent & condition = Combined	**1.689**	**1.458**	**1.158**	**0.247**

displaying a frequency trade-off in their use of HND vs. LND words. However, there was a significant main effect of condition, such that participants in the COMBINED condition were more likely *overall* to use the HND words than participants in the COMPREHENSION condition. A full summary of model coefficients is given in Table 4.1. The model's predictions for each combination of condition and object frequency are shown in Figure 4.14.

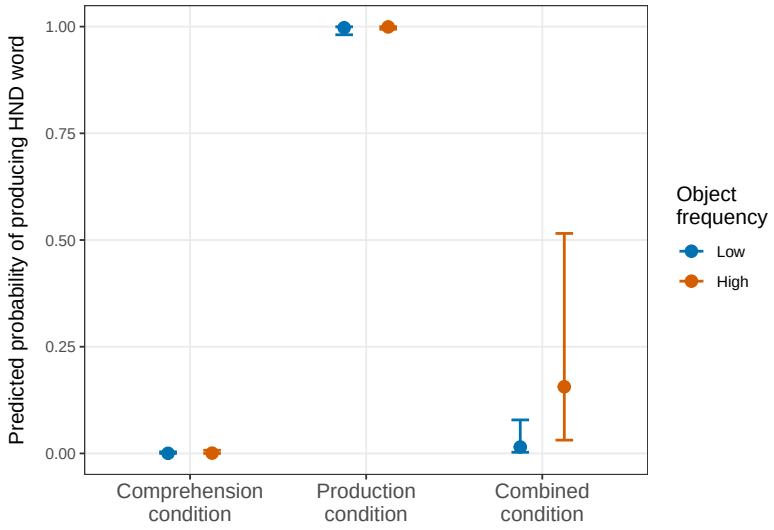

Figure 4.14: Model predictions for each combination of condition and object frequency, generated using the `ggeffects` package (Lüdecke 2018). Points represent the predicted probability of producing an HND word; error bars represent the 95% confidence interval around this value. Although the model predicts that participants in the critical COMBINED condition were numerically more likely to produce an HND word for the high-frequency object than the low-frequency object, this interaction between condition and frequency was not statistically significant (see Table 4.1).

4.4.2.2 Exploratory analysis

Figure 4.13 suggests that when only one aspect of the communicative task was difficult, most participants took the same approach to mitigating this difficulty: data points are

strongly clustered in the bottom-left and top-right corners in the COMPREHENSION and PRODUCTION conditions respectively. By contrast, when both aspects of the task were difficult, it is less clear that participants were converging on a single optimal solution: data points are more widely scattered around the plot in the COMBINED condition. In particular, there are a number of points towards the centre of the plot (on at least one axis) in this condition, representing pairs who appear to be probability matching to the input by using the HND and LND words approximately 50% of the time each (for at least one object). However, this method of visualisation disguises some underlying differences between the two members of the pair. Specifically, while it is possible that a pair at the centre of this plot could consist of two participants probability matching to the input, it is equally possible that these points represent pairs where one participant is only using the HND words and the other is only using the LND words. Indeed, if we plot individual participants instead of collapsing across pairs, we can see that the data tends to move away from the centre and towards the corners (Figure 4.15).

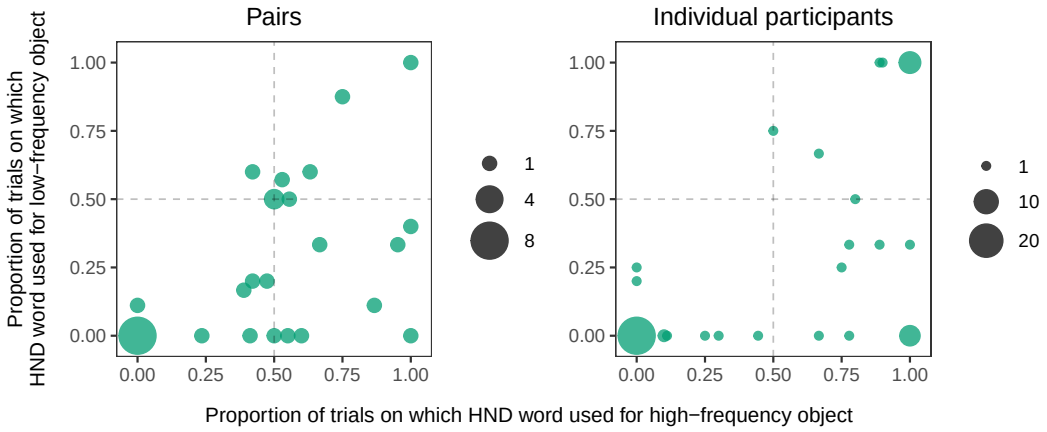

Figure 4.15: By-pair (left) vs. by-participant (right) data for the COMBINED condition. Although it appears that a number of pairs are producing HND and LND words with roughly equal frequency, it is clear that individual participants are at least somewhat consistent in their choice of word. This suggests that pairs towards the centre of the left-hand panel have not converged on a shared language; rather, these pairs probably consist of one participant who is mostly using the HND words for both objects and one who is mostly using the LND words for both objects.

To further explore this trend, we calculated a convergence score for each pair by comparing the languages produced by each member of the pair. Each participant's output language can be fully described by a 2-dimensional vector (*HF, LF*) where *HF* is the proportion of trials on which the participant used the HND word for the high-

frequency object and *LF* is the proportion of trials on which they used the HND word for the low-frequency object. For example, the vector $(1, 0)$ captures a language showing the expected frequency trade-off (i.e. in the bottom-right corner of the plot). The *divergence* between two members of a pair is given by the Euclidean distance e between their output languages. The maximum possible Euclidean distance between two n-dimensional vectors is equal to $\sqrt{n}$ when the input values are bounded between 0 and 1. Therefore, the *convergence* between two members of a pair is given by $\sqrt{2} - e$. Figure 4.16 shows the distribution of convergence scores by condition. We fit a linear regression model to this data, predicting convergence score as a function of experimental condition (treatment-coded with the COMPREHENSION condition as the reference level). The model reveals that within-pair convergence was significantly lower in the COMBINED condition ($\beta = -0.407$, $SE = 0.107$, $t = -3.804$, $p < 0.001$), while there was no significant difference between the COMPREHENSION and PRODUCTION conditions ($\beta = -0.073$, $SE = 0.107$, $t = -0.682$, $p = 0.497$).

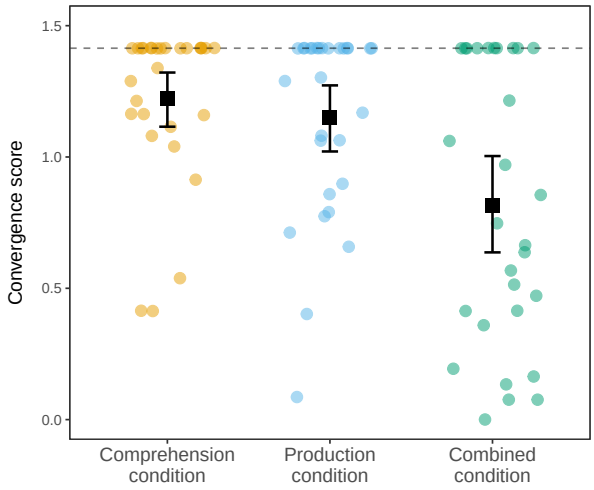

Figure 4.16: Convergence scores by condition. The dashed line indicates the maximum possible score, which is achieved when both members of a pair produce exactly the same output language. Each coloured point represents an individual pair. Black points represent the mean over all pairs in that condition; error bars represent bootstrapped 95% confidence intervals over the mean. Convergence scores are similarly high in the COMPREHENSION and PRODUCTION conditions, but significantly lower in the COMBINED condition.

Since pairs in the COMBINED condition are often failing to converge on a shared language, we might also expect accuracy on Matcher trials to be lower in this condition. Figure 4.17 shows how often the Matcher successfully selected the target object in each condition, depending on the object's frequency and the word used to label it. We fit a logistic mixed effects model to this data, predicting accuracy as a function of exper-

imental condition (treatment-coded with the COMPREHENSION condition as the reference level), word type (treatment-coded with LND as the reference level), object frequency (treatment-coded with low-frequency as the reference level), and all two-way and three-way interactions between them. The model also included by-participant random intercepts, but failed to converge with random slopes for object frequency or nested random intercepts by-participant and by-pair. There was no main effect of being in the COMBINED condition ($\beta = -0.389$, $SE = 1.120$, $t = -0.347$, $p = 0.728$). However, the model yielded a significant three-way interaction between condition, frequency and word type, such that the probability of a correct response was higher in the COMBINED condition when the target object was high-frequency and labelled with the HND word ($\beta = 4.136$, $SE = 1.607$, $t = 2.574$, $p < 0.05$).

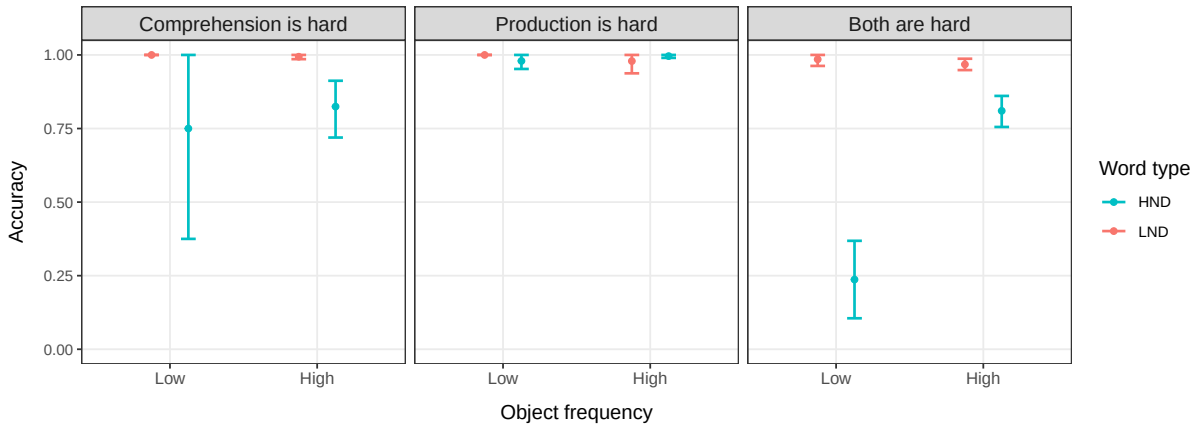

Figure 4.17: Accuracy on Matcher trials by condition, object frequency and word type. Accuracy is high across the board for LND words, which are always unambiguous. Accuracy for HND words depends both on condition and object frequency: participants in the COMBINED condition are significantly more likely to successfully infer the intended meaning of these words when they are used to label the high-frequency object than when they are used to label the low-frequency object, suggesting that participants in this condition may have some expectations of a natural-language-like frequency trade-off when interpreting ambiguous signals.

This three-way interaction could indicate that participants had some expectations of a natural-language-like frequency trade-off in comprehension (even if this was not borne out in their productions). Specifically, participants were relatively successful at inferring their partner's intended meaning when an HND word was used to label the high-frequency object, even though the information provided by the word form alone could equally point to either object. Conversely, participants were very unlikely to infer that their partner was referring to the low-frequency object when they used an HND word. However, it is difficult to determine whether this discrepancy only

arises in the COMBINED condition because participants in this condition understand that there are pressures in favour of both HND and LND words and therefore form different expectations about how their partner might be behaving, or because this is the only condition where both word types are used frequently enough to observe a difference between them. In other words, it may be that accuracy for HND words only appears to be similar across the two object frequencies in the COMPREHENSION condition because these words are hardly ever used for either object[13]. If this is the case, then accuracy for HND words in the COMBINED condition may simply reflect a strategy of guessing meanings proportional to their frequency when the signal is ambiguous (i.e. guess the high-frequency meaning 75% of the time and the low-frequency meaning 25% of the time).

4.4.3 Experiment discussion

In our experiment, we found that language users were easily able to adapt their lexical choices for efficient communication when *only* production was difficult or *only* comprehension was difficult. However, the picture was less clear when both of these pressures were present. Some participants converged on the efficient natural-language-like solution: mapping easy-to-produce but potentially ambiguous words to frequent objects and harder-to-produce but easily distinguishable words to infrequent objects. However, other participants apparently prioritised one pressure over the other, either by using only the unambiguous LND words despite their cost in production, or by using only the easily accessible HND words despite their cost in comprehension. Nonetheless, as in our model, the lexicons that emerged when production and comprehension pressures were in competition represented an intermediate state between the extreme outcomes observed when only one of these pressures was at play, at least in terms of the *overall* likelihood of producing an HND word.

Notably, this experiment was designed as a relatively close replication of Kanwal et al. (2017). Although the exact production and comprehension pressures we simulate are not identical, the net effect of these pressures was very similar: LND words (like

[13] Accuracy in the PRODUCTION condition is, unsurprisingly, at ceiling across the board, since the clean transmission channel in this condition ensures that all words are unambiguous.

long words in Kanwal et al.) took longer to produce, and HND words (like short words in Kanwal et al.) were ambiguous in communication. Despite these parallels, we do not replicate the frequency trade-off that arose in Kanwal et al.'s COMBINED condition. In considering why our findings did not robustly bear out our predictions, it is worth laying out what might have led to this discrepancy.

Certainly, the two experiments do differ in a number of important ways. Firstly, the input languages are quite unalike. The two objects in Kanwal et al.'s experiment shared a short name ("zop") which was derived by clipping their unique long names ("zopekil" and "zopudon"). In this way, there was a clear relationship between an object's alternative names, and the ambiguity of the short name was a property of the lexicon that was evident throughout the experiment, including during training. Conversely, the two names for each object in our experiment were clearly unrelated, and while the HND words were very similar to each other, there was no outright ambiguity in the lexicon: the ambiguity only arose during communication as a side-effect of noisy transmission. It may therefore be the case that participants in Kanwal et al. were starting to form ideas about how they would deal with the ambiguity earlier in the experiment, whereas participants in our experiment had insufficient time to explore different strategies once they realised that the HND words were functionally ambiguous. In fact, it is possible that participants in our experiment didn't even realise that the HND words *were* ambiguous for their partner; anecdotally, a handful of participants reported on the debrief questionnaire that their partner was only sending one-letter responses, suggesting that not all participants understood that the noisy transmission was symmetrical and their partner had the same kind of comprehension difficulty as themselves. This is an inherently different situation from the one in Kanwal et al., where participants knew exactly how much information the different labels provided for for their partner. Furthermore, it is likely that participants have more explicit awareness and experience of abbreviating frequent words (e.g. "information" → "info") than they do of preferentially selecting between synonyms to maximise ease of pronunciation, and may be bringing this experience to bear when considering how to solve the task.

Secondly, the manipulation of production effort in Kanwal et al. was perhaps more transparent than our keyboard task: the time for which participants had to click and

hold to send a longer word in the former was effectively dead time, whereas participants in our experiment were still engaged in the task whilst forming LND words, even if it did take longer. Although our manipulation clearly works in the sense that participants in the PRODUCTION condition strongly favoured the easier-to-form HND words, it could still be the case that it is too subtle when a competing pressure is present. This may also be exacerbated by the fact that the pressure for accuracy probably feels inherently stronger for participants than the pressure for speed: Prolific participants are highly motivated to complete tasks "correctly" to avoid having their submissions rejected. We tried another version of the experiment which attempted to address these first two points (reported in Appendix 4.A), but the effect of frequency was not obviously stronger in this follow-up; the most noteable change in participants' behaviour was simply an increased preference in favour of the HND words *overall*.

Finally, long words in Kanwal et al. remained consistently arduous throughout the experiment, since they always took a fixed number of seconds to transmit. On the other hand, participants in our experiment may have been able to improve at the keyboard task, thereby reducing the cost to produce LND words over time (relative to the cost for their partner by *not* producing them). However, we think this is unlikely to account for much of the variance between the two experiments since the letters required to form LND words changed position on every trial, so the only thing participants could really learn that would help them produce these words on subsequent trials is that they could ignore the centre of the keyboard (which should have become obvious almost immediately).

Nonetheless, our experiment does provide further evidence that neither production pressures nor comprehension pressures *alone* give rise to the kind of organisational structure we see in real lexicons, in line with Kanwal et al.'s results regarding Zipf's Law of Abbreviation and with the results of our computational model when it comes to word similarity. Furthermore, to the extent that there are subtle tendencies towards a natural-language-like frequency trade-off when both pressures are present, we would expect these to be amplified through transmission to successive generations of participants (Reali & Griffiths 2009; K. Smith & Wonnacott 2010; Thompson et al. 2016).

4.5 General discussion

In this paper, we investigated how pressures operating during individual episodes of communication might give rise to an emergent structural property of language, whereby lexicons tend to be more phonetically clustered than required by their phonotactics, especially for high-frequency items.

In an exemplar-based computational model, we showed that clustering emerges under competition between production-side pressures for word similarity and comprehension-side pressures for discriminability. The lexicons that arise from this competition are neither as clustered nor as disperse as they possibly could be, although there is some variance in the exact details of how the two pressures are balanced depending on the strength of the comprehender-side pressure for distinctiveness and, to a lesser extent, frequency. With only one communicative pressure at work, the resulting lexicons very clearly fall at one extreme or the other. Specifically, when producibility is the only pressure, the outcome of repeated communication is a lexicon that is extremely easy to produce but communicatively degenerate, in that all words sound almost exactly the same. On the other hand, when comprehensibility is the only pressure, lexicons are maximally expressive in that all words are very distinct, but arduous from a production perspective due to the lack of shared sound sequences across words.

In a communication experiment using an artificial language, we showed that, when ease of production is the only pressure shaping participant behaviour, a strong preference emerges in favour of words from a high-density neighbourhood, while when ease of comprehension is the only pressure, the opposite preference (in favour of words from low-density neighbourhood) emerges. Extrapolating these preferences to an imagined wider lexicon, it is clear that our experiment makes the same predictions as our model: production pressures alone would be expected to give rise to a highly clustered lexicon, while comprehension pressures alone would lead to a highly disperse lexicon. As in the model, an intermediate state emerges when these pressures are in competition. Specifically, one neighbourhood does not completely win out over the other in this scenario; rather, words from both neighbourhoods have their place. However, it is not clear that selection between words from the different neighbourhoods is modulated by frequency.

Putting these two pieces together, our results demonstrate that mechanisms operating during individual episodes of communication can shape the structure of the lexicon. Crucially, we show that evolving lexicons balance the influence of competing pressures that pull in different directions. However, with respect to the role of frequency, our results are less clear: frequency effects were subtle in our model, and do not emerge robustly in our experiment. Clearly, it is not possible to make precise predictions from natural language data about what effect sizes we would expect in such highly simplified, simulated lexicons. However, it is worth noting that the relationship between frequency and clustering in real languages is not necessarily a strong one; in fact, it is specifically described as a "weak tendency" by Frauenfelder et al. (1993). Correlations between frequency and different measures of clustering in Mahowald et al. (2018) were generally small, with Pearson's r values deemed as statistically significant starting at 0.08 and rarely exceeding 0.3. The relationship between frequency and clustering may also be stronger for word beginnings than endings (King & Wedel 2020), or for content words over function words (Frauenfelder et al. 1993), factors not considered here. Therefore, we would suggest that the subtlety of the frequency effect across our model and experiment may be exactly as expected.

One criticism that might be levelled at our study is that the extreme outcomes that emerge under the influence of a single communicative pressure paint a highly unrealistic picture of the cognitive biases that shape language. As pointed out by Wasow et al. (2005), if our notion of "production effort" includes the effort required to clarify what was intended for a confused receiver, then effort would clearly not be minimised by a degenerate language (with only one word for every meaning). However, in the limit, a bias to re-use sound sequences across words points to exactly such a language, and we would argue that, all else being equal, producers would want their language to conform to this bias. It is exactly because producers have communicative goals that all else is *not* equal, and a compromise position has to emerge. Similarly, it is clearly true that, as comprehenders, we can happily cope with some amount of noise in the linguistic signal, because there are plenty of other ways to extract an interlocutor's intended meaning — from contextual cues in the environment to the many multimodal features of language like co-speech gesture and facial expression. Even so, if all language users cared about was maximising comprehensibility, there would certainly be no harm in

having lexicons be as disperse as their phonotactics would allow. It is precisely because comprehensibility is *not* the only thing language users need to worry about that we do not see such lexicons in the real world. Whilst acknowledging that these counterfactual either-or situations do not represent real language use, it is still useful to examine their consequences in isolation; by doing so, we can verify that the phenomena we are trying to explain do in fact result from a trade-off between competing pressures, and cannot be more simply explained by one pressure or the other.

Natural language lexicons, as in the critical conditions of our model and experiment, are under pressure to adapt to several competing forces. The way in which they achieve an optimal balance between these pressures is clearly not simple, and depends on several factors. For example, biases can vary in strength: in our model, one source of variation was captured by the Receiver's γ parameter (Section 4.3.1.3), but there are no doubt others in the real world, such as differences in articulatory or auditory apparatus that might make certain sound sequences more or less difficult to pronounce for certain individuals (e.g. Franken et al. 2017). In our experiment, a variety of individual differences may have pushed different participants to arrive at different solutions to the task; for example, more risk averse participants may have been less willing to sacrifice accuracy for the sake of speed (Carver & White 1994). Nonetheless, the lexicons that emerge under competing pressures are, in some sense, *efficient* (Gibson et al. 2019; Jaeger & Tily 2011): words are just distinctive "enough" whilst still being as easy to produce "as possible" (where "enough" and "as possible" are defined with reference to a specific communicative or cognitive context). Optimising for producibility inevitably means introducing some ambiguity, but as pointed out by Piantadosi et al. (2012), ambiguity is actually a hallmark of an efficient communication system since it allows for the reuse of words and sounds that are more easily produced, and doesn't impede communication as long as there are other ways for the comprehender to overcome the ambiguity. In our experiment, for example, participants could overcome the ambiguity of the HND words during Matcher trials either by adopting a very simple heuristic of probability matching their guesses to the relative frequencies of meanings in the world (since words are, *a priori*, more likely to refer to things we talk about more), or by establishing a shared code with their partner that would allow them to use probabilistic information from previous interactions to inform future ones.

While our study provides further evidence for the role of competing communicative pressures in driving language efficiency, our simulation of the pressures acting on language is undoubtedly a simplification in a number of ways. Mostly notably, our experiment *simulates* the pressures involved in language use, rather than relying on them to emerge at scale in the lab. Most obviously, typing is not language production in the usual sense, and naturalistic comprehension is not the same as image selection. Replicating this study in a more ecologically valid setting (i.e. with oral production and auditory comprehension tasks) is a logical next step for a few reasons. First, allowing pressures to emerge naturally could, in principle, provide more compelling evidence for a causal link between individual-level behaviour and population-level language trends like phonetic clustering. Second, there may be specific aspects of production effort that are not well-simulated by anything other than oral production. However, it seems likely that the difficulty associated with these tasks would still need to be artificially inflated — for example, through the use of highly phonotactically complex words, or environmental noise on transmission — to observe, in a brief experiment, the kinds of effects that otherwise accumulate only over much larger timescales. The benefit of our design is that it allows us to easily manipulate task difficulty in a way that affects all participants roughly equally and does not depend on, for example, prior experience with pronouncing certain sounds, or auditory acuity. By doing so, we can get an idea of how small and potentially noisy effects at an individual-level might accumulate into large effects at a population-level (Kirby et al. 2007).

The present work also does not account for every possible mechanism that could play a role in shaping this aspect of lexicon structure. For example, it is possible that clustering emerges more strongly from new words entering the lexicon than from changes to or selection between existing words. Such a mechanism could also go some way to explaining the frequency effects we see in natural languages: if high-frequency words are a stronger attractor for the form of new words than low-frequency words, new coinages would tend to increase connectivity more in high-frequency components of the lexicon (see Dautriche et al. 2017a for a similar suggestion). Future work should investigate how different kinds of lexical evolution — from coinage to sound change and, ultimately, obsolescence — might differentially drive changes in the network properties of the lexicon.

Furthermore, neither our model nor our experiment account for the role of learning biases in shaping linguistic systems (Christiansen & Chater 2008; Culbertson 2012; Griffiths et al. 2008; Kalish et al. 2007; Kirby et al. 2008, 2014; K. Smith et al. 2003b). There are several reasons to think that learning might play a role in driving increased clustering. For one, lexicons built from a smaller inventory of sound sequences are more compressible (Ferrer-i-Cancho et al. 2013), a property which reduces storage demands (Storkel & Maekawa 2005) and allows languages to pass more easily through the bottleneck imposed by repeated transmission to naive individuals (Kirby et al. 2015). Moreover, infants and children show clear preferences for words composed of the highest-frequency sound sequences in their target language (Altvater-Mackensen & Mani 2013; Jusczyk et al. 1994; Ngon et al. 2013) and generally acquire such words earlier (Coady & Aslin 2004; Gonzalez-Gomez et al. 2013; Storkel 2004). Since early-acquired words are also known to be more stably represented within a community's language (Monaghan 2014), we might expect these developmental effects to show up in evolution. However, a learning-based account does not straightforwardly point to a clustering advantage (see e.g. Dautriche et al. 2015; Jones & Brandt 2020; Storkel & Lee 2011; Storkel et al. 2006; Swingley & Aslin 2007).

Finally, lexicons are not, contrary to the dominant view of "design features" (Hockett 1960), entirely arbitrary. Rather, languages are rife with sound symbolism and other systematic associations between form and meaning (Bergen 2004; Blasi et al. 2016; Cuskley & Kirby 2013; Dautriche et al. 2017b; Dingemanse et al. 2015; Monaghan et al. 2007, 2014; Tamariz 2008). A detailed account of the role of semantics is missing from our study, since there is no level of analysis below the atomic meaning (e.g. we do not consider the meaning "lightbulb" to have any features that might be shared across other meanings, such as being man-made or having to do with electricity). However, while correlations between semantic similarity and wordform similarity are significantly higher than would be expected by chance, effect sizes are generally very small (Dautriche et al. 2017b; Monaghan et al. 2014), so this is unlikely to be the main driver of phonetic clustering in natural language lexicons. Another source of non-arbitrariness is shared etymology: words that come from the same historical root may consequently sound similar in their modern form (Klein 1971). We do not take into account any such structure in our models since we use randomly-generated

lexicons as the input to the agents. However, we would argue that if the phonetic clustering that resulted from shared etymology was detrimental for communication, it could be selected out through cultural evolution; the fact that natural language lexicons are observably more clustered than they could be suggests that this is not the case. Nonetheless, future work could look to incorporate notions of semantic and historic relatedness as a more conservative test of our hypotheses. Our model could also be adapted to test a variety of different starting conditions.

4.6 Conclusion

Corpus data shows that natural language lexicons are more phonetically clustered than would be expected, even accounting for phonotactic rules, morphology and sound symbolism. This study provides the first evidence that this organisational property of the lexicon can arise as a result of mechanisms operating at the level of individual language users and individual communication episodes. Specifically, we show that emergent lexicon structure balances the influence of competing functional pressures: a pressure for distinctiveness arising from comprehension, and a pressure for reuse of forms arising from production. When only one of these pressures is present, the lexicons that emerge exhibit extreme levels of clustering or dispersion unlike those seen in natural languages. This study adds to a growing body of evidence showing that, through a process of cultural evolution, languages are optimised for efficient communication.

4.A Follow-up experiment

As discussed in Section 4.4.3, there were a number of differences between the design of our experiment and the one it was modelled after (Kanwal et al. 2017). In particular, we felt that our manipulation of production effort may have been too subtle to push participants towards an efficient solution in the presence of a competing pressure for accuracy. We also wondered whether the unclear relationship between an object's two alternative names may have changed participants' representation of the language in a way that could influence their behaviour during communication. We therefore ran a follow-up experiment which attempted to address these two concerns, while maintaining the general design whereby words from the high-density neighbourhood were easier to produce but functionally ambiguous, while words from the low-density neighbourhood were harder to produce but easily distinguishable. The changes are summarised in Figure 4.18 and described below.

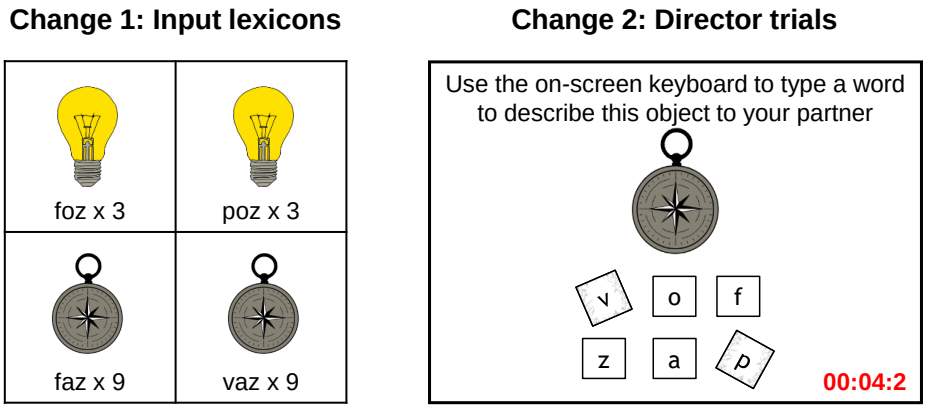

Figure 4.18: Summary of design changes in the follow-up experiment. Input lexicons were designed such that the HND words were clearly variants of the LND words, rather than completely different words (left). Director trials used an on-screen keyboard in which the keys required to form an LND word were faulty — indicated by their cracked texture and wonky placement — and sometimes produced an incorrect letter (right).

4.A.1 Materials

The meaning space consisted of the same two objects in the same frequency distribution as in the first experiment. The language consisted of four artificial CVC words: "foz" /fɑz/ and "faz" /fæz/ (the HND words) and "poz" /pɑz/ and "vaz" /væz/ (the LND words). Each LND word in this lexicon has a corresponding HND word (with

which it shares the final two phonemes) which is derived by a known process of sound change: /p/ → /f/ (e.g. Foulkes 1997) and devoicing as in /v/ → /f/ (e.g. van de Velde et al. 1996).

4.A.2 Procedure

The procedure was identical as in the first experiment, except for the design of the difficult Director trials. On these trials, as before, the Director was presented with both word options for the target object and asked to use an on-screen keyboard to type one of the words. However, the keyboard in this experiment contained only letters that were part of the artificial language, and all buttons were the same size and appeared in the same position from trial-to-trial (the configuration was randomised for each participant). Instead, the two keys required to make an LND word ("p" and "v") were wonky (a random angle of ±10, ±15 or ±20 degrees was chosen for each button on each trial), and had a cracked texture around the edge. At the start of each trial, a random integer between 1 and 3 was generated, representing the total number of times either of these keys would need to be pressed before the correct letter would appear; other times, a random letter that wasn't part of the artificial language would appear. Every time one of these keys produced an incorrect letter, participants would need to press an "undo" button to get rid of that letter before trying again. Participants were told that some of the buttons were faulty and might need to be pressed a few times. As before, this design was intended to simulate the observation that less frequently-used phonemes are more error prone; however, we hoped that this manipulation would make the LND words more costly from participants' perspective than in the first experiment.

4.A.3 Participants and exclusions

Due to financial constraints, we were only able to run the critical COMBINED condition in this follow-up experiment. We used Prolific to recruit 72 participants who had not taken part in the first experiment. The experiment took around 25 minutes to complete in full (median time = 22:44) for which participants were paid £4.25. One participant was prevented from proceeding to the communication game due to low accuracy on

the pre-test and paid a reduced rate of £2. 13 participants started but failed to complete the interaction phase and were paid a variable rate depending on how far they had got through the experiment. Two participants (one pair) completed the communication game and were paid the full rate, but their data was excluded from analysis because their completion time was more than 3 standard deviations above the median. After all exclusions and dropouts, we were left with 28 pairs: a total of 56 individual participants.

4.A.4 Results

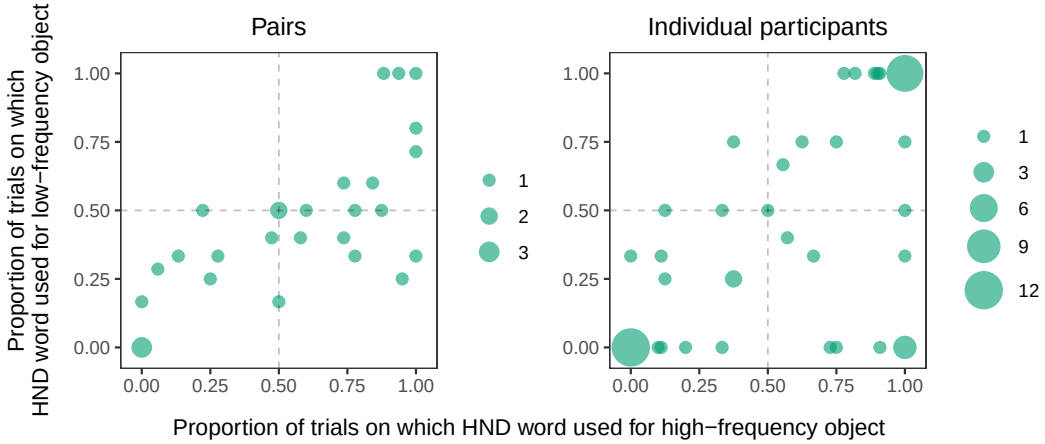

Figure 4.19: Proportion of trials on which the HND word was used for the high-frequency object vs. the proportion of trials on which it was used for the low-frequency object, by-pair (left) and by-participant (right). As in the first experiment, individual participants are more strongly clustered in the corners than pairs, suggesting that not all pairs are converging on a shared language. Also as in the first experiment, a range of behaviours are represented, and it is not clear that a natural-language-like frequency trade-off (bottom right quadrant) is the most common strategy.

Figure 4.19 shows the proportion of Director trials on which the HND word was used for the high and low-frequency objects. As in the first experiment, a range of strategies are represented, and it is not clear that most participants are converging on the predicted frequency trade-off. We fit a reduced version of the model described in Section 4.4.2.1; since we only ran one condition in this follow-up experiment, there is no longer a fixed effect of condition, nor an interaction between condition and frequency. The model had by-participant random intercepts and random slopes for object frequency, but failed to converge with the nested by-pair random effects structure used in Section 4.4.2.1. Model predictions are shown in Figure 4.20. The model reveals a

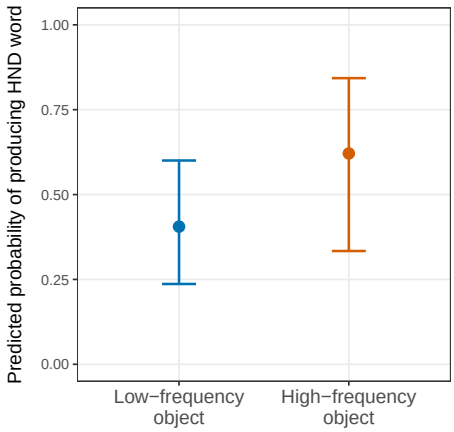

Figure 4.20: Model predictions generated using the `ggeffects` package (Lüdecke 2018). The model predicts that participants were more likely to produce an HND word for the high-frequency object than for the low-frequency object.

significant main effect of frequency, such that participants were more likely to use the HND word to label the high-frequency object ($\beta = 0.877$, $SE = 0.392$, $t = 2.237$, $p < 0.05$). This result follows straightforwardly from the fact that there are many more participants below than above the diagonal in Figure 4.19 i.e. for participants who showed *any* effect of frequency, it was generally the predicted one. In other words, very few participants adopted an anti-efficient strategy of using the difficult-to-produce LND word for the high-frequency object and the the easy-to-produce HND word for the low-frequency object.

However, if we consider the two experiments as a whole, it seems that the key difference between them is not in the strength of the frequency effect. We pooled the data from the COMBINED condition of the first experiment with the data from this follow-up experiment, and fit a mixed effects logistic regression model predicting HND word use as a function of object frequency, experiment, and their interaction. Again, the model had by-participant random intercepts and random slopes for object frequency, but failed to converge with a nested by-pair random effects structure. A full summary of model coefficients is given in Table 4.2. The model reveals no overall effect of frequency, despite the significant effect of frequency when considering the follow-up experiment in isolation. However, there is also no interaction between frequency and experiment; that is, there is no evidence that either experiment showed a clearer effect of frequency. Crucially, the model does show a significant main effect of experiment, such that the *overall* probability of producing an HND word was higher in the follow-

Table 4.2: Summary of fixed effects for a logistic mixed effects model with HND word use as the binary dependent variable and by-participant random effects for object frequency. The main experiment reported in Section 4.4 is labelled as 1a; the follow-up experiment is labelled as 1b. Coefficient estimates are on the log-odds scale.

	β	SE	z	p
intercept (object = infrequent, experiment = 1a)	-3.039	0.707	-4.300	<0.001
object = frequent	1.537	0.799	1.923	0.054
experiment = 1b	2.546	0.851	2.993	<0.01
object = frequent & experiment = 1b	-0.452	0.944	-0.479	0.632

up experiment. In other words, our changes to the experimental design succeeded in making the LND words more costly for participants to produce, but not in such a way that made the predicted frequency trade-off emerge more robustly. Convergence between the two members of a pair (i.e. the extent to which they settled on a shared language) also did not improve in the follow-up experiment (Figure 4.21).

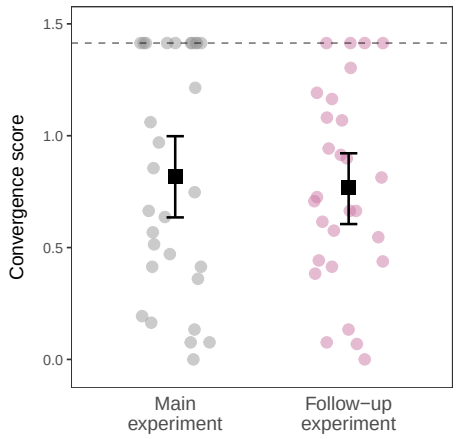

Figure 4.21: Convergence scores for the COMBINED condition of the main experiment (left) and the follow-up experiment (right). Convergence is very similar between the two experiments.

Overall, the results of this follow-up experiment provide further evidence that, insofar as there is a relationship between frequency and clustering, it may be more subtle than the relationship between frequency and word length probed by Kanwal et al. (2017)'s experiment.

4.B Oral production: A pilot study

As I described in Chapter 1 (Section 1.3.1), a reasonable criticism of this thesis is that I am interested in language production, but my experiments are really only proxies for production. I therefore wanted to try running a more naturalistic experiment, with auditory stimuli and spoken production. In what follows, I report a pilot study which was intended as a replication of the critical COMBINED condition in the main experiment reported in this chapter: both production and comprehension are hard, and I am interested in whether participants will exhibit a frequency trade-off in their productions which balances between these two pressures.

4.B.1 Materials

The meaning space consisted of two objects from the NOUN database (Horst & Hout 2014) with high novelty scores[14]: only 6% of participants in the norming study indicated that they had seen either of these objects before. I used different stimuli for this experiment because I was concerned that, given the freedom of oral production, participants might be tempted to use English words during the test phase; I chose objects with low nameability to mitigate against this possibility. The objects appeared in the same frequency distribution as in the other experiments. The language consisted of four artificial words, each comprised of four CV syllables (delineated in the transcriptions by hyphens): /bæ-bæ-du-peɪ/ and /bæ-bæ-du-keɪ] (the HND words) and /si-ʃeɪ-ʃoʊ-su/ and /feɪ-θɔɪ-θu-fi/ (the LND words). The LND words were designed to mimic tongue twisters — an ABBA pattern of syllables with phonetically similar onsets — and therefore cause difficulty in pronunciation (Acheson & MacDonald 2009; Croot et al. 2010; Wilshire 1999). The HND words were designed to be easy to pronounce, and I am confident they are because they are based on a made-up alien name (*Babadoolish*) that we use in a workshop with primary school children (ages 7-11).

[14] In the NOUN catalogue, objects 2013 and 2025.

4.B.2 Procedure

Training On each training trial, an object was presented on screen for 1500ms while the audio file of the appropriate word played once. Participants were instructed to repeat the word they had just heard and then click a 'next' button to advance to the next trial; they were able to replay the audio file as many times as they wanted to, and the research assistant made sure they repeated the word correctly.

Pre-test The pre-test phase was the same as in the other experiments except that the artificial words were presented auditorily, not orthographically. Audio files played once automatically; after this, participants were able to replay the audio as many times as they wanted to.

Simulated interaction Participants who successfully completed the pre-test were told that they were going to play a communication game with a partner who was in another room, and that they would be randomly assigned a role to play throughout the game: Director or Matcher. However, this was a cover story: the interaction phase was fully simulated. Participants always played as the Director, and the computer played as the Matcher. To make the cover story a little more convincing, participants were put into a simulated waiting room before the communication game began: a random wait time between 10 and 60 seconds was generated for each participant. Once this time had elapsed, participants were informed that their partner was ready, and were asked to show the research assistant their screen so she could start the game. The research assistant gave the following instructions: "I'll be making sure that you're using valid words from the new language. I'll use an external keyboard to start the next round once you've said a valid word. The only thing you need to do each round is click the microphone button to start recording, and click it again to stop recording."

On each trial, participants saw an object and were asked to name it for their partner. As explained by the research assistant, they clicked an on-screen microphone button to start and stop audio recording. Once the participant stopped recording, the research assistant pressed a key to record the type of response the participant had given; the key mapping is given in Table 4.3.

Table 4.3: Key mapping used by the research assistant to record participants' responses during the simulated communication game. A "valid" response was a correct word for the target object, pronounced correctly. An "attempted" response was an incorrect pronunciation of a correct word. The "other" category was intended as a catch-all for a variety of possible behaviours e.g. using a word that belonged to the other object, using an English word, asking for a reminder of the available words, or making an attempt at an artificial word that did not obviously come from the language the participant was trained on.

Key	Meaning
f	Valid HND word
q	Attempted HND word
j	Valid LND word
p	Attempted LND word
spacebar	Other invalid response

If the research assistant pressed any key except "f" or "j", she explained to the participant that they had made a mistake and needed to try again, and then restarted the trial. I instructed the research assistant that she could give participants a hint if they didn't understand what they had done wrong or if they forgot one of the words, but to use her judgement and try and give as little information as possible. Luckily, she reported that participants always realised what they had done wrong, and she never had to give any more specific information than "that wasn't quite right". Once the research assistant indicated that the participant had said a valid word, the participant saw a screen informing them that they needed to wait for their partner to make a response; a random wait time between 1 and 5 seconds was generated on each trial. Participants received feedback on whether their "partner" had selected the correct picture.

The computer played the role of a Matcher for whom the HND words were functionally ambiguous but the LND words were entirely unambiguous (as in the critical COMBINED condition of the main experiment). To simulate the Matcher's behaviour, I used a probabilistic model. If the participant used an LND word, the computer responded correctly with probability 0.95. If the participant used an HND word, the computer sampled a meaning from an unordered collection, which I'll call C. At the start of the interaction phase, C contained 9 copies of the frequent object, and 3 copies of the infrequent object: the input frequencies. However, as the game proceeded, the computer kept track of the frequency with which the participant used an HND word to refer to the two objects. Specifically, every time the participant used an HND word,

a copy of the target meaning for that round was added to *C*. This is illustrated in Figure 4.22.

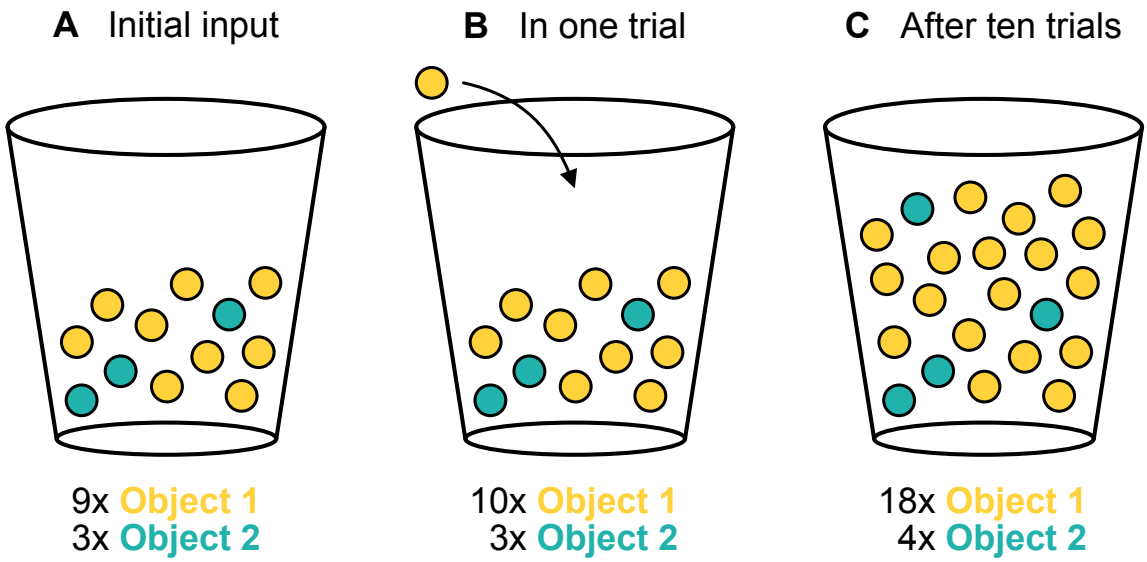

Figure 4.22: Schematic of the probabilistic model playing the role of Matcher. Initially, upon hearing an ambiguous HND word, the model would just be sampling from the input frequencies for the two objects, giving it a 75% chance of choosing the high-frequency object (A). On the first trial where the participant uses an HND word, the model adds another copy of the target meaning for that trial to its collection (B); this gives the model a slightly higher probability of choosing that object on future trials. If the participant is consistent in which object they prefer to label with an HND word, the model will become increasingly likely to sample this object after more and more trials (C).

4.B.3 Participants and exclusions

Due to time constraints, the research assistant was only able to recruit seven participants in time for inclusion in this thesis. Participants were recruited from the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh using the SONA system. The majority of participants in this system are first year undergraduate students in psychology. They are compensated with 0.5 course credits per study. All participants reported that English was their first language and that they had no known language disorders. They attended the lab in person and completed the experiment on a laptop with built-in microphone whilst seated across from the research assistant. The experiment was approved by the PPLS Ethics Committee at the University of Edinburgh (ref. 6-2425/4). One participant was prevented from proceeding to the communication game due to low accuracy on the pre-test (but still received their credit).

4.B.4 Preliminary results

Since my sample size is so small at the time of writing, I will not present any inferential statistics here. Instead, I will just provide some exploratory analysis of the data collected so far.

Figure 4.23 shows the proportion of trials on which the HND word was used for the high and low-frequency objects. The general picture seems quite similar to the other experiments: a range of behaviours are represented, and there is no clear trend towards the expected frequency trade-off. Overall, participants used the HND word for the high-frequency object 44.6% of the time and for the low-frequency object 43.8% of the time.

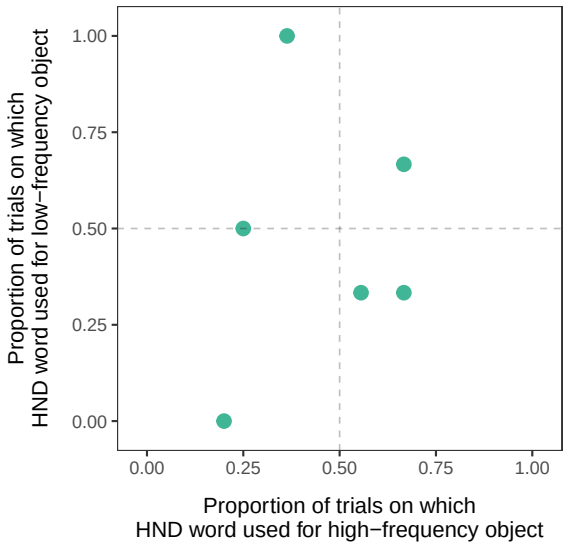

Figure 4.23: Proportion of trials on which the HND word was used for the high-frequency object vs. the proportion of trials on which it was used for the low-frequency object, by participant (there is no by-pair data for this experiment since participants played with the computer rather than with another human). As in the other experiments, a range of behaviours are represented, and it is not clear that a natural-language-like frequency trade-off (bottom right quadrant) is the most common strategy.

In total, participants completed 179 production trials. Of these, 35 (almost 20%) were recorded as being invalid in some way. The error rate differed by object frequency: 15 of 51 trials for the low-frequency object (29.4%) were recorded as invalid, compared to 20 of 128 trials for the high-frequency object (15.6%). Individual participants made anywhere between 1 and 9 errors throughout the test phase (mean = 5.8, median = 6.5). Figure 4.24 shows the proportion of invalid trials that fell into each of the three categories: an attempt at an LND word, an attempt at an HND word,

or any other kind of invalid response. By far the most common type of invalid response for both object frequencies was an attempt at an LND word, which suggests that these words were — as intended — harder to pronounce than their HND counterparts. Reassuringly, very few responses fell into the "other" category, which suggests that participants were engaged in the task and making a genuine attempt to communicate accurately in the artificial language.

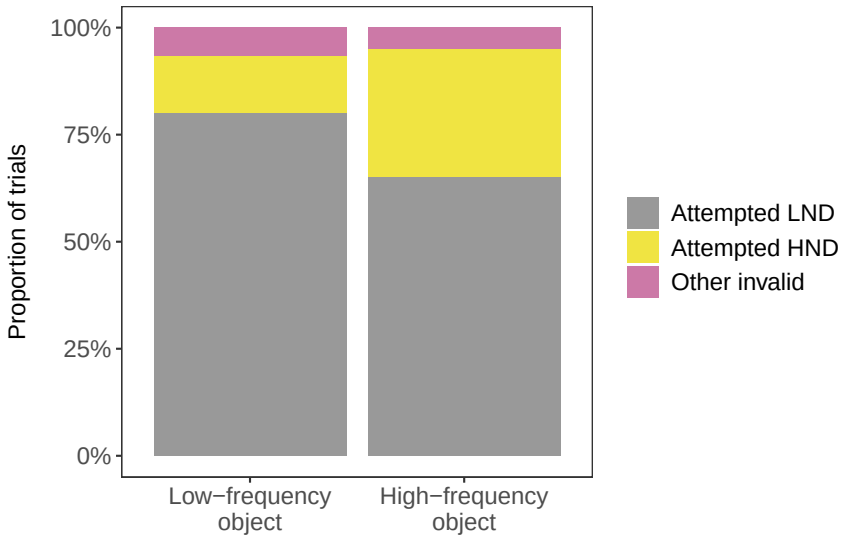

Figure 4.24: Proportion of invalid production trials that fell into each category. For both object frequencies, the most common error type by far was an incorrect pronunciation of an LND word — as expected, since these words were designed to be harder to pronounce. Invalid responses in the "other" category were very rare for both object frequencies, suggesting that participants were generally trying to complete the task as intended.

4.B.5 Discussion

Since I can't draw any real conclusions from this pilot data, I'll just spend this section reflecting on some aspects of the experiment design and also the experience of collecting data in-person vs. online.

Obviously, one big difference between this experiment and the two others reported in this chapter is that it was not genuinely communicative. Anecdotally, the research assistant reported that most participants didn't seem to realise their partner wasn't real (they were debriefed at the end of the experiment), so this may not have made much difference to their behaviour in the end[15]. However, it is worth thinking about how

[15] And in fact, a handful of participants in the other experiments made comments suggesting that they didn't believe their partner *was* real.

human-like the computer was as a Matcher in this task. Imagine a participant who *always* used the HND word to label the high-frequency object, and *never* used this word to label the low-frequency object. By the end of the test phase, the computer would have 27 copies of the high-frequency object in its collection, and the original 3 copies of the low-frequency object: a 90% chance of choosing the high-frequency object on hearing an ambiguous HND word. For sure, this shows they have learned *something* about the participant's behaviour, but have they learned they same kind of information as a human would have done? Intuitively, I wouldn't imagine a human taking this long to realise that their partner's behaviour was entirely consistent, and to become similarly consistent in their guesses. If participants in this experiment thought that their "partner" was still making lots of mistakes when they used an HND word, even when they were using it consistently for only one object, this might have incentivised them to switch to the LND words for both objects. The appeal of the model I used for the Matcher is that it makes very few assumptions: it's the most neutral possible implementation I could have used. However, it arguably does retain too much probability mass on a behaviour the participant may not actually exhibit by remembering the input frequencies forever; perhaps a more realistic model would be one that weighted more recent evidence more highly (or forgot older evidence entirely).

The use of a simulated Matcher was only possible in the first place because there was a research assistant in the room, manually coding participants' speech in real-time into a representation that the computer could understand. On the plus side, this meant that we didn't need to worry about trying to get pairs of participants into the lab at the same time (SONA participants are notoriously unreliable!), it vastly minimised the time and effort I had to spend on the analysis, and it meant I didn't need to think of a clever way to make the HND words ambiguous during communication (e.g. by adding white noise to participants' audio recordings on-the-fly). However, the downside is that in-person data collection is considerably slower than online; in fact, this is precisely why I only have a pilot-sized sample here.

More generally, I think we also have to wonder whether participants would have behaved differently if they were not being so closely observed. One of my major concerns with using an oral production task has always been data quality, and having the research assistant in the room was intended to alleviate this concern. The low num-

ber of responses labelled "other invalid response" is reassuring on this front, but I'm not convinced it would replicate online: without the supervision of a research assistant, what would be to stop participants from using English words or otherwise deviating from the artificial language? Data quality in my other experiments is highly controlled, since I can reliably (and programatically) prevent participants from submitting an invalid response. This means that all the rows that appear in my dataframe are analysable, and I'm not shelling out participant payments for unusable data.

Overall, if I were to summarise this brief experience of trying to collect more naturalistic data, I would say it was a challenge: both logistically, and in terms of nailing down the design. Potentially a fruitful avenue for future research would be a typed production task, which might be a happy middle ground. Such a task could be more easily administered online to speed up data collection, and allows for more easily automated verification and analysis techniques.

4.C Reanalysis of data from Kanwal et al. (2017)

When I was conducting the analysis of the experiment reported in Section 4.4 and realised that participants in the COMBINED condition were often not converging on a shared language, I thought it would be worth having another look at the data from the equivalent condition in Kanwal et al. (2017) to see if the picture was similar there. I suspected that this wouldn't be the case, since the model coefficients reported in their Table 1 were from a by-participant model, which clearly showed the expected frequency trade-off. However, their Figure 3 (the equivalent of my Figure 4.13) only showed the data by-pair, so it wasn't obvious whether there might be any interesting differences between pair behaviour and individual participant behaviour.

Figure 4.25 shows the comparison between by-pair and by-participant data. It looks like there are a couple of individual participants in the top-right quadrant (using the short word for both objects), which was not obvious from the by-pair plot. However, on the whole, the two plots look much more similar than in my Figure 4.15: the bulk of the data is in the bottom-right on both, in line with the predicted frequency trade-off.

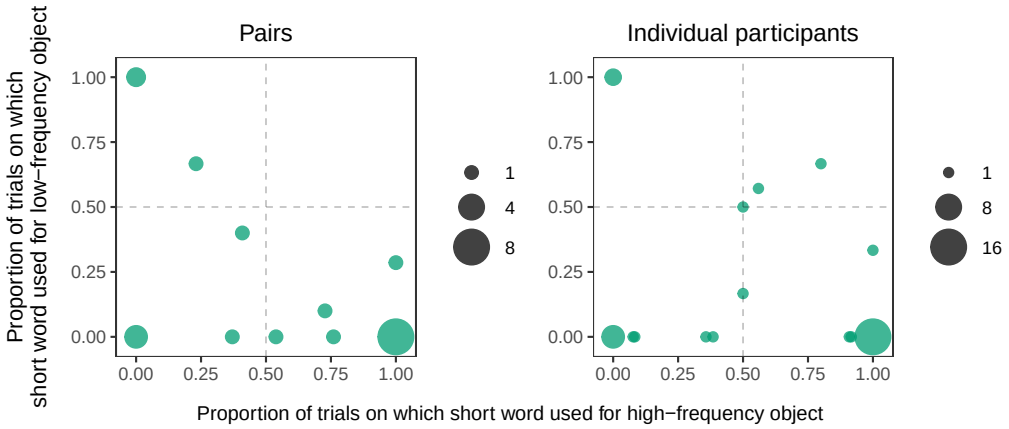

Figure 4.25: By-pair (left) vs. by-participant (right) data for the COMBINED condition from Kanwal et al. (2017). On the whole, the data looks very similar whether we average over the two members of a pair or not.

I also calculated convergence scores for this condition and the ACCURACY condition. I am treating the latter as the equivalent of my COMPREHENSION condition: participants in this condition needed to communicate accurately, which, all else being equal, would favour the unambiguous long words, and there was no countervailing

pressure from the production-side in favour of the short words. Figure 4.26 shows the distribution of convergence scores by condition. Unlike in my experiment, the scores are very similar between the two conditions: specifically, they are similarly *high*. That is, participants in both conditions were generally converging on a shared language. A linear regression model predicting convergence scores as a function of experimental condition (treatment-coded with the ACCURACY condition as the reference level) revealed no significant difference between conditions ($\beta = -0.066$, $SE = 0.078$, $t = -0.848$, $p = 0.402$).

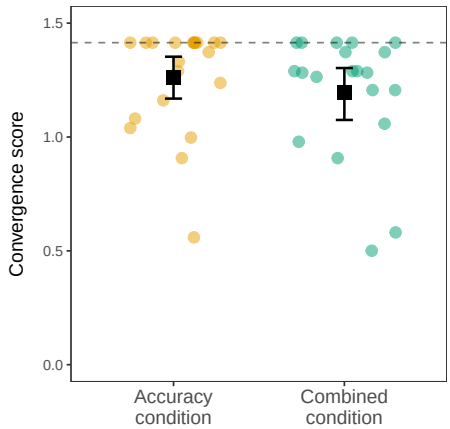

Figure 4.26: Convergence scores by condition in Kanwal et al. (2017). Convergence is similarly high (on average, near ceiling) across conditions, indicating that participants were generally converging on a shared mapping between words and objects.

Overall then, it seems that the convergence problem was specific to my experiment: participants in Kanwal et al. (2017) were managing to establish a shared language.

4.D Reanalysis of data from Kirby et al. (2008, 2015)

Very early on in this project, I wondered whether it might be possible to use existing data to test my hypothesis that languages would become more phonetically clustered as they evolved. Specifically, I turned to two classic iterated learning studies: Kirby et al. (2008) and Kirby et al. (2015). In these experiments, participants were trained on artificial languages describing structured meaning spaces i.e. the meanings varied on multiple dimensions, like shape, colour or fill pattern. However, the input languages were completely *un*structured: randomly generated words were randomly distributed across the meanings. Once participants had been exposed to these languages, they then had to try and reproduce them. In Kirby et al. (2008), participants did this testing phase by themselves. In Kirby et al. (2015), they were paired up to play a communication game where they took turns as Speaker and Hearer: the Speaker had to type a word to describe a picture, and the Hearer had to guess which picture the Speaker was describing. The set of labels produced by one participant (or pair of participants) were then passed on to a new generation of participants, and the whole process repeated.

In both studies, as the generations proceed, the languages change from the initial input. The *way* in which the languages change provides a window on the processes by which natural languages might change over time. Specifically, the key findings as reported in these papers were as follows. In single-participant chains (Kirby et al. 2008), there is a pressure for learnability being imposed at each new generation, but no pressure for expressivity since the languages are not being used for communication. In the absence of any intervention (Experiment 1), these conditions give rise to increasingly underspecified languages, where the same word is used to convey multiple meanings. In the limit, languages can become *degenerate*: one word for every meaning. In a second experiment reported in Kirby et al. (2008), an anti-homonymy filter was imposed to prevent this kind of underspecification; under these conditions, the languages evolved to be *compositional*, with smaller pieces of linguistic material ("morphemes") reused across words. When the languages were used for communication (Kirby et al. 2015), they remained expressive, but the way in which they achieved this differed depending on the transmission process. In one condition (CLOSED GROUP), each pair was retrained on their own productions every generation: there were no naive learners.

In this condition, the languages retained their initial, unstructured (*holistic)* form e.g. while a checkered leaf might be described as "gakho", a spotty leaf would be described with a completely unrelated word, like "wuwele". In the other condition (CHAIN), the languages produced by each pair were transmitted to a new pair of participants, again giving rise to a pressure for learnability. In this condition — as in Experiment 2 from Kirby et al. (2008) — the languages became *compositional*: different parts of the words systematically mapped to different parts of the meaning e.g. "mega-wawa" for checkered leaf, "mega-wuwu" for spotty leaf, and "ege-wuwu" for spotty bean.

I wondered whether it might be possible to observe increased sound similarity between the words of the evolving languages, independent of these other effects. To try and get at this question, I calculated the same average pairwise edit distance measure I used in Section 4.3. However, this time, I normalised it word for length: the normalised edit distance between two words is the raw distance divided by the length of the longer word. I didn't need to do this normalisation in the model because all words were the same length, but it was an important step here since words were of varying lengths, and not accounting for this could give rise to spurious conclusions: if average word length decreases, edit distance will necessarily also decrease, and if average word length increases, edit distance will necessarily also increase. Normalising edit distance puts it on a scale from 0 to 1, where 0 means that two words are identical, and 1 means they are as different as they could possibly be.

Figure 4.27 shows the change in normalised average pairwise edit distance over generations for each transmission chain in the two studies. In Kirby et al. (2008), there is a clear decrease in edit distance across the two experiments, such that words are becoming more similar. This is entirely unsurprising given the summary I gave earlier of what happens under pressure for learnability alone: the languages are becoming increasingly underspecified, and using the same word (or parts of words) for multiple meanings inevitably gives rise to increased clustering. For the data from the CHAIN condition in Kirby et al. (2015), the picture is less clear, but it still looks like the overall trend might be in the downward direction. Again, this is as expected: compositional languages are more clustered than holistic ones, because they re-use morphemes across different words. So neither of these results are particularly informative. I was most interested to see what would happen in the CLOSED GROUP condition, since the lan-

guages that emerged in this condition had no underspecification or compositionality that could drive clustering. However, it doesn't look like there's much going on here either: in some chains, there's a modest decrease in edit distance from Generation 0 (the initial input) to Generation 1 (the first output language produced by a pair of participants). However, the lines look pretty flat for most chains after this point, and in hindsight, this too was completely predictable: Figure 4 in Kirby et al. (2015) clearly shows that the languages in this condition really didn't change *at all* after the first "generation". That is, participants in this condition very quickly became familiar with the language and stopped making mistakes.

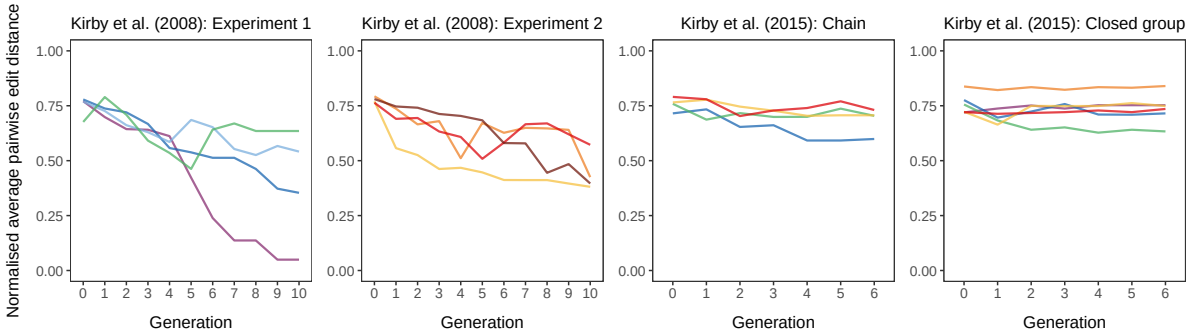

Figure 4.27: Change in normalised average pairwise edit distance over generations in the experiments reported in Kirby et al. (2008) and Kirby et al. (2015). Each coloured line represents a different transmission chain. Edit distance inevitably decreases when languages become more underspecified *or* more compositional (leftmost three panels). When participants are retrained on their own productions and the language is never transmitted to naive learners (rightmost panel), edit distance doesn't change very much after the first "generation", since participants become very familiar with the language and stop making mistakes.

I also realised that there was not necessarily much room for the languages in these experiments to become more clustered while remaining fully expressive, just because of the way the input was designed. In both experiments, the initial set of labels was generated by concatenating between two and four CV syllables. In each chain, these syllables were sampled from a set of nine, which itself was randomly selected from a larger set comprising all possible combinations of eight consonants {g, h, k, l, m, n, p, w} and five vowels {a, e, i, o, u}. This relatively small space of letters and letter combinations means that the initial lexicons were already somewhat clustered: Generation 0 in each of the plots in Figure 4.27 is nowhere near the ceiling of 1. And although participants were not forced to use only letters/syllables that were present in the initial input, they generally did stick to using letters they had seen themselves: on average,

each generation of participants only introduced 1.39 (SD = 3.36) letters that were not present in their input in Kirby et al. (2008), and 1.60 (SD = 3.80) in Kirby et al. (2015). So it may have been quite difficult for words to become any more similar to each other whilst maintaining essential distinctions, given the size of the letter inventory available to participants.

Overall then, it turned out that the existing data wasn't sufficient to answer the questions I was interested in. And from this very equivocal start, the rest of this chapter was born!

4.E Additional model analysis

In the paper that forms the main body of this chapter, I included results from three versions of the computational model: one with only production pressures, one with only comprehension pressures, and one with both. Eagle-eyed readers may have been wondering what happens when *neither* of these pressures are at play. To get at this question, I ran another version of the model where both the *retrieval bias* and *error bias* parameters were switched off: all exemplars had equal probability of being retrieved for production, and errors simply replaced one random segment with another random segment. I also set *context size* to minimal, such that there was no inference on the Receiver's part[16]. I consider this combination of settings to be the null model: the only source of variation is random drift, and there is no selection of particular types of signals because communication is always successful.

I ran the null model both with and without the frequency manipulation described in Section 4.3.3.1: results are shown in Figures 4.28 and 4.29 respectively. For ease of comparison, I've included the plots from the paper as well. In both cases, where there are no communicative pressures that might favour different variants, lexicons remain in their maximally-disperse starting state. This suggests that drift alone — through random sampling of exemplars and random errors in production — does not inevitably give rise to more clustered lexicons, and the non-random mutation that arises from production biases is an essential component.

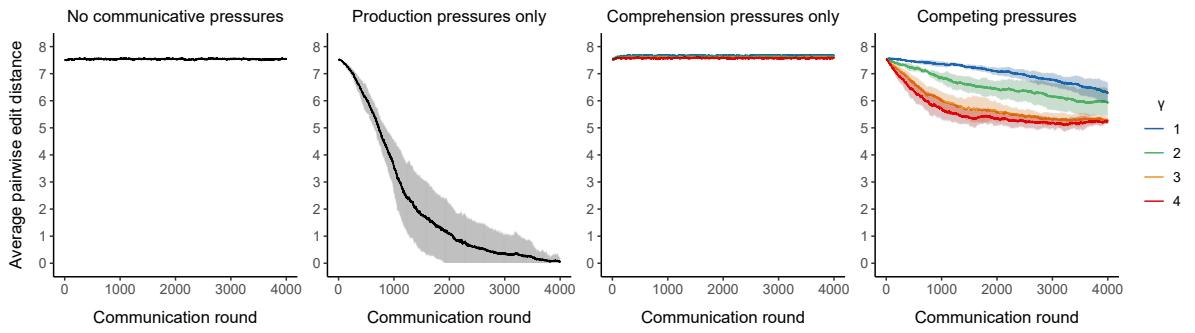

Figure 4.28: The same model results presented in Figure 4.7, now including the null model with no communicative pressures (left). Drift alone does not pull lexicons out of their starting state.

[16] An equivalent way of achieving this is to remove communication from the equation entirely, by just having one agent produce signals and update their own internal representation repeatedly.

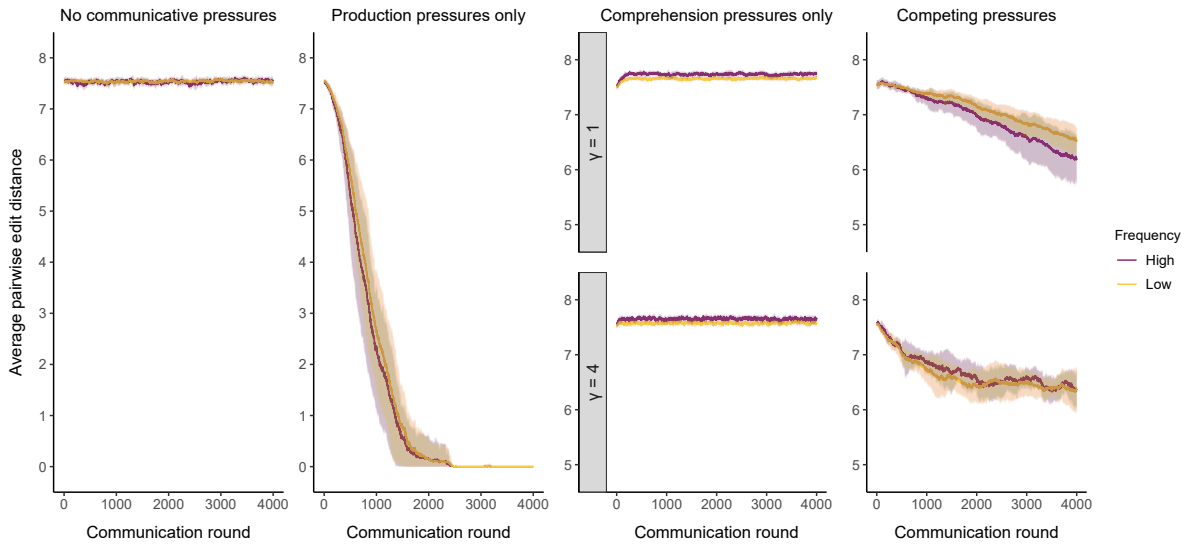

Figure 4.29: The same model results presented in Figure 4.8, now including the null model with no communicative pressures (left). Drift alone does not pull lexicons out of their starting state.

I also wondered what would happen if I started the model from a different kind of input lexicon, one that already exhibited a moderate degree of clustering. I predicted that the outcome at the end of iteration would be essentially the same as when starting from a random lexicon: production pressures would lead to maximal clustering, comprehension pressures would lead to maximal dispersion, and the competition between them would result in an intermediate state. However, when starting from an already-clustered lexicon, these predictions point to a different direction of change in the latter two conditions: if words are starting off very similar, then the influence of comprehension pressures should be making them *less* similar over time.

To start the model off with a more clustered lexicon, I modified the initialisation process described in Section 4.3.1.2. Specifically, instead of generating 20 words (one per meaning category), I generated just *two* words. Half of the meaning categories were mapped to one word, and the other half to the other word. I then used these words to seed the process of exemplar creation for each category (in the same way as described in Section 4.3.1.2). This approach meant that the starting lexicons were, on average, about half as clustered as they could possibly be. In other words, any two randomly chosen words had a 50% chance of being identical, and a 50% chance of being (probably very) different. Concretely, average pairwise edit distance was bounded between 0 and 8 (the word length); in the original models, it was generally close to 8 at the start,

whereas in these models, it was generally around 4.

Figure 4.30 shows the results for this model configuration in the same four conditions as above. Overall, the results for the non-null conditions are basically as expected: words become maximally similar in the production-only condition, maximally dissimilar in the comprehension-only condition, and somewhere in between with both of these pressures at work.

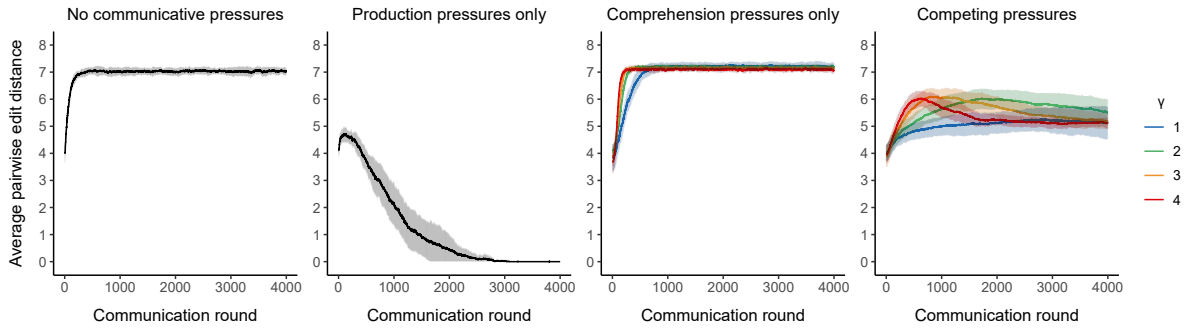

Figure 4.30: Results of the model in four conditions (including the null model: left) when starting from a more clustered input lexicon (using only two seed words, instead of 20). As before, production pressures alone give rise to maximally clustered lexicons, comprehension pressures alone give rise to maximally disperse lexicons, and an intermediate state emerges under their combined influence.

However, some of the details seem a bit strange at first glance. Firstly, in the right-hand two plots, it looks like lexicons are taking longer to become more dispersed when the Receiver's γ parameter is low, which is exactly when the pressure for dispersion is *strongest*: with lower values of γ, the Receiver relies on a greater degree of dispersion to be able to tell words apart. In hindsight though, this result makes sense: selection (or replication) of signals only happens after a successful communication episode, and communication will be successful less often when the Receiver's γ parameter is low. When there's no selection, the lexicon doesn't change. So in effect, when communication is noisier, the rate of change is necessarily slower.

The second aspect of these results I found a bit surprising is the fact that the null model looks *exactly the same* as the comprehension-only model: lexicons very rapidly hit the ceiling in both. Of course, this was also the case for the previous version of the model, but I had imagined that things would be different in this version: I expected the lexicons in the comprehension-only model to become more disperse, and those in the null model to just fluctuate around the starting state. Again, in hindsight, this

was clearly not the right prediction. To quote from a supervision meeting where we discussed these results: "If you go on a drunken walk, you're not very likely to end up where you started." Drift is always going to make things more random, and there are more ways to be different than there are to be similar. However, I do still feel a little uncomfortable about the fact that drift and selection end up looking indistinguishable, and I haven't yet thought of a good way to reconcile this issue.

4.F Details of the rule-based phonotactics baseline

The content of this appendix was provided by Juan Guerrero Montero; it describes the phonotactic rules implemented in his code to generate artificial words conforming to English phonotactic constraints (used in the corpus study in Section 4.1).

4.F.1 Onsets

The following onsets were permitted with no restrictions on the following nucleus:

- Empty onset.
- All single consonants except /ŋ/: /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /j/, /w/, /r/, /l/, /f/, /θ/, /s/, /ʧ/, /ʃ/, /h/, /v/, /ð/, /z/, /ʒ/, /ʤ/.
- Stop plus approximant other than /j/ (note that /pw/ is attested but /bw/ is not): /pl/, /bl/, /kl/, /gl/, /pr/, /br/, /tr/, /dr/, /kr/, /gr/, /tw/, /dw/, /gw/, /kw/, /pw/.
- Voiceless fricative or /v/ plus approximant other than /j/: /fl/, /sl/, /θw/, /ʃl/, /fr/, /θr/, /ʃr/, /hw/, /sw/, /θw/, /vw/.
- /s/ plus voiceless stop, or plus nasal other than /ŋ/, or plus voiceless non-sibilant fricative: /sp/, /sk/, /st/, /sm/, /sn/, /sf/, /sθ/.
- /s/ plus voiceless stop or voiceless fricative, plus approximant other than /j/: /spl/, /skl/, /spr/, /str/, /skr/, /skw/, /sfr/.

The following onsets were permitted only if followed by /uː/ or /ʊr/:

- Consonants other than /r/ or /w/ plus /j/ (note that many clusters allowed in RP, such as /nj/, are not in General American): /pj/, /bj/, /kj/, /gj/, /mj/, /fj/, /vj/, /hj/.
- /s/ plus voiceless stop or nasal, plus /j/: /spj/, /stj/, /skj/, /smj/.

4.F.2 Nuclei

Allowed nuclei are given in Table 4.4.

Table 4.4: Restrictions on nuclei. Each column indicates whether the given vowel is allowed in the described position. Some vowels can be considered as different phonemes when they appear before an /r/ coda, but for simplicity they are represented with the same symbol here. Schwa is unrestricted and can appear in any position in the word. *Only if followed by /r/.

Vowel	End of word	Before another vowel	Before /r/ coda	After /j/
/ei/	Yes	Yes	No	No
/i/	Yes	Yes	No	No
/ai/	Yes	Yes	No	No
/ou/	Yes	Yes	No	No
/u/	Yes	Yes	No	Yes
/æ/	No	No	No	Yes
/ɛ/	No	No	Yes	No
/ɪ/	No	No	Yes	No
/ɑ/	No	No	Yes	No
/ʊ/	No	No	Yes	Yes*
/ɔ/	Yes	No	Yes	No
/ə/	Yes	No	Yes	No
/oi/	Yes	Yes	No	No
/au/	Yes	Yes	No	No

4.F.3 Codas

The following codas were permitted with no restrictions on the preceding nucleus:

- Empty coda (restricts preceding nucleus if at the end of a word or next to an empty onset).
- All single consonants except /h/, /w/ or /j/: /b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /ŋ/, /r/, /l/, /f/, /θ/, /s/, /ʧ/, /ʃ/, /v/, /ð/, /z/, /ʒ/, /ʤ/.
- Lateral approximant plus stop, affricate, fricative, or nasal: /lp/, /lb/, /lt/, /ld/, /lk/, /lʧ/, /lʤ/, /lf/, /lv/, /lθ/, /lð/, /ls/, /lʃ/, /lm/, /ln/.
- Nasal plus stop, affricate, or fricative: /mp/, /nt/, /nd/, /nʧ/, /nʤ/, /ŋk/, /mt/, /md/, /ŋd/, /mf/, /mz/, /mθ/, /nθ/, /ns/, /nz/, /ŋθ/.
- Fricative plus stop of the same voicing: /ft/, /sp/, /st/, /sk/, /ʃt/, /θt/, /zd/, /ðd/.

- Two voiceless fricatives or two voiceless stops: /fθ/, /pt/, /kt/.
- Stop plus fricative: /fθ/, /pt/, /kt/, /pθ/, /ps/, /tθ/, /ts/, /dθ/, /dz/, /ks/.
- Three-consonant clusters: /lmd/, /lpt/, /lps/, /lfθ/, /lts/, /lst/, /lkt/, /lks/, /mpt/, /mps/, /nts/, /ntθ/, /ŋkt/, /ŋks/, /ksθ/, /kst/.

The following codas placed restrictions on the preceding nucleus, as detailed in Table 4.4:

- Empty coda restricts nuclei if at the end of a word or next to an empty onset.
- /r/ and clusters starting with it: /rp/, /rb/, /rt/, /rd/, /rk/, /rg/, /rʧ/, /rʤ/, /rf/, /rv/, /rθ/, /rð/, /rs/, /rz/, /rʃ/, /rm/, /rn/, /rl/, /rmd/, /rmθ/, /rpt/, /rps/, /rnd/, /rts/, /rst/, /rld/, /rkt/.

4.F.4 Intersyllabic restrictions

Contiguous syllables are checked to make sure that the same sound and voiceless/voiced versions of the same sound are not in contact. Other very specific restrictions were not implemented under the assumption that they would not affect the statistics of the generated sample with respect to measures like edit distance. These unimplemented restrictions include the fact that /ʊ/ is rare in syllable-initial position, and sequences of /s/ + C1 + V + C1 (where C1 is a consonant other than /t/ and V is a short vowel) are virtually non-existent.

Chapter 5

General discussion

Stop discovering things!

Professor Simon Kirby

In complete disregard of my supervisor's advice, this thesis *has* made several contributions to our understanding of how language structure emerges under the influence of production pressures. In this final chapter, I will provide a summary of where we've been[1], before suggesting some ideas for where we could go next.

5.1 Summary of contributions

5.1.1 Regularity in language as shaped by production *and* learning

First, in Chapter 2, I investigated the mechanisms that underly *regularisation*, a well-documented process whereby languages become less variable (on some dimension) over time.

In an artificial language learning experiment, I taught participants a language which indicated plurality on nouns with two different markers, and manipulated working memory demands during either the initial learning phase or the testing phase (when

[1] As I'm writing, my husband keeps reminding me of this advice (which may originally be from Aristotle, or from an anonymous preacher in the early 1900s, we'll never know): "Tell 'em what you're gonna tell 'em; tell 'em; tell 'em what you told 'em". So this will be the part where I tell you what I told you.

participants had to produce phrases in the artificial language themselves). I observed that the latter manipulation — memory load during production — led to an overall reduction in variability, such that one of the plural markers was boosted at the expense of the other. This result held even when the choice between the two markers in the input was predictable — some nouns used one marker, some used the other. I also found that, when no such patterns existed in the input, participants introduced them in their own productions; this behaviour was not driven by the same memory load mechanism, but rather seemed to stem from participants' difficulty in encoding a completely random mapping between nouns and plural markers.

In a computational "urn" model (inspired by Spike et al. 2017), I tested one hypothesis about the source of the memory load effect: that when working memory is more limited, recently produced variants will be disproportionately likely to be retrieved again — essentially, a process of self-priming. By exploring a wide parameter space on this model, I formalised a set of assumptions — both about the population make-up, and about the nature of the priming process — that could give rise to the same pattern of results as the human participants.

Overall, the main contribution of this chapter is to demonstrate that systematic rules and regularities in language can arise from an interaction between two factors: constraints on working memory during online production, and a learning bias against randomness.

5.1.2 Learning morphology through production *or* comprehension

In Chapter 3, I collaborated with Elizabeth Pankratz to address a shared curiosity, namely: can a production task boost learning of a difficult morphological rule, compared to a comprehension task?

In two artificial language learning experiments, we taught participants a language that indicated thematic role both through its word order, and through nominal case marking morphology. We designed the input such that the morphological rule was, to a certain extent, hidden: participants received no direct evidence that the case marking suffixes carried a grammatical meaning, and were not just meaningless syllables that

reoccurred across different nouns by pure coincidence. After a period of passive exposure, we had participants practise the language with either a more active, production-like task (assembling syllables into sentences), or a more passive comprehension task (reading sentences and choosing a matching picture). We then tested which rule(s) participants had learned by asking them to judge new sentences that unambiguously followed or did not follow a case marking grammar.

We hoped that participants in the production group — who had a chance to actively manipulate the smaller parts within the words — would pick up more strongly on the pattern of reoccurring syllables than the comprehension group, and would realise that it would be a suspicious coincidence for these syllables to reoccur for no reason. However, across both experiments, we found no evidence that the type of practice task affected participants' generalisations about the language: everyone learned the word order rule, and almost no one learned the case marking rule. The only benefit we found for our production task was that it seemed to boost participants' familiarity with the sentences they had been trained on: we excluded fewer participants from the production group for low accuracy on these items.

We suspect that the null result with respect to our main hypothesis can be attributed to some shortcomings in our experimental design. For example, our use of written stimuli would have discouraged participants from segmenting below the word level, since whitespace in the orthography inherently makes a distinction between words but not between parts of words.

Overall, the results of this chapter are inconclusive. However, we still believe the question is an interesting one, and would benefit from further experimentation. In our discussion, we suggested a range of follow-up studies which we believe would improve upon our design, and, in doing so, might shed more light on the role of language production for learning of morphological rules.

5.1.3 Word similarity as a trade-off: production *vs.* comprehension

Finally, in Chapter 4, I explored how competing communicative pressures might shape patterns of word similarity in the lexicon.

I started with a small corpus study in which I compared real words of English against a range of random and phonotactically-controlled baselines. This study replicated previous work showing that, even when controlling for the effects of productive morphology and word length, real words are surprisingly similar to each other — more similar than they need to be given the available phonotactic space (Dautriche et al. 2017a). At first glance, this property — which I refer to as *phonetic clustering* — is a surprising one for a lexicon to have, given that increased phonetic similarity makes it harder to tell words apart.

To investigate what mechanisms could give rise to this seemingly surprising level of clustering, I developed an agent-based computational model of communication. I built in a variety of biases that have been observed in psycholinguistic experiments. Specifically, in production, words that are more similar to other words are more easily retrieved and more accurately pronounced, while in comprehension, words that are more *different* from other words are more easily recognised. I initiated the model with a random lexicon, and let pairs of agents communicate repeatedly to see how this lexicon would evolve. I found that natural-language-like lexicons emerged as a compromise between competing pressures for and against word similarity. With only one of these pressures at work, the outcomes were implausibly extreme: production pressures alone resulted in maximally clustered lexicons (one word for every meaning), while comprehension pressures alone resulted in maximally disperse lexicons (no shared sound sequences across words). I also showed that, in principle, these same communicative pressures could generate a subtle frequency trade-off, whereby more frequent words become more tightly clustered and less frequent words remain more distinctive.

Finally, in a series of experiments, I tested whether human participants would arrive at this same frequency trade-off when communicating in an artificial language. In the main experiment, I trained participants on a language in which different objects appeared with different frequencies, and were labelled by words with different phonological properties. I then paired participants up to play a communication game with this language, during which I simulated production and comprehension pressures that made more similar words easier to produce but harder to tell apart. My hypothesis was that participants would adapt their lexical choices to facilitate effi-

cient communication depending on which pressures were active. And indeed, as in the model, when only one of the pressures was present, participants took the obvious approach: always using the easily-producible words when only production was hard, and always using the easily-understandable words when only comprehension was hard. However, the picture was less clear when production and comprehension pressures were in competition. Some participants converged on the optimal solution — using an easily-producible word for the frequent object, and a more distinctive word for the infrequent object — but many other kinds of behaviour were also represented. In a couple of follow-up studies, I tried some small tweaks to the experimental design to see whether the expected frequency trade-off would emerge more robustly under different conditions, but the results remain inconclusive about the effect of frequency on the emergence of clustering. On the whole though, the experiments accord with the model, in that the "average" behaviour under the influence of competing pressures was less extreme than under just one pressure, representing some kind of balance between the needs of producers and comprehenders.

Overall, the main contribution of this chapter is to illustrate how biases operating in individual language users and individual communication episodes can, through a process of cultural evolution, give rise to emergent structure in the lexicon. I propose that the patterns of word similarity I studied in this chapter can be seen as a generalisation of the Law of Abbreviation (Zipf 1949): lexicons are organised such that, whether by their length or by their phonological make-up, words are as easy to pronounce as possible, and as easy to recognise as necessary.

5.1.4 Overview: revisiting the three key ideas

In Chapter 1, I set out three key ideas that guided the work in this thesis: (1) that producing language in real-time comes with a host of cognitive and motor challenges, (2) that the challenges posed by production have significant implications for language structure, and (3) that language is ultimately shaped by the interplay between production pressures and other functional pressures arising from learning and comprehension.

Across the projects summarised above, I have addressed at least three challenges as-

sociated with production: it requires us to retrieve the correct items from memory and inhibit competitors (Chapter 2), assemble these items into ordered sequences (Chapter 3), and hope that our motor articulators can pronounce them without error (Chapter 4).

I have shown that, on multiple timescales, these challenges have implications for language structure: production difficulty drives regularisation of morphosyntactic variation (Chapter 2), potentially leads to more robust learning of linguistic rules (although not necessarily different *kinds* of learning: Chapter 3), and gives rise to a surprising level of phonetic similarity between words (Chapter 4).

Finally, I have explored several ways in which language is shaped by competing pressures. The clearest case of this was in Chapter 4, where production and comprehension pressures were in direct competition, and the tug-of-war between them resulted in an intermediate state between the extremes that would arise under just one or the other. In Chapter 2, the pressures I investigated were not *competing* as such, but rather coexisting: both learning biases and production difficulty led to greater regularity, just on different dimensions. It's also important to point out that, while I do believe the demands of online production can drive behaviour which is not a transparent reflection of what has been learned, learning and production are not really in an either/or relationship: both feed in to each other. Finally, in Chapter 3, the competition I imagined was between a shallower representation of the language's grammar formed through passive comprehension, and a deeper representation formed through active production; I still think there's good reason to believe this is the way real language learning works, but unfortunately, our results didn't enable us to identify any such difference for the specific case study we looked at.

5.2 "Future research could..."

I have highlighted potential avenues for future research within each of the content chapters, so I don't want to dwell for too long here on what's left for those mysterious "future researchers"[2]. However, in rounding up these three projects, I do just want to

[2]It's me. I am the future researcher.

briefly discuss some methodological extensions that could apply across my work more generally.

5.2.1 More naturalistic production tasks

First and foremost, the obvious extension to all of the work presented in this thesis is to adapt the experiments to use more naturalistic production tasks — speaking, signing, writing or typing could all plausibly tap into some important mechanisms that are missing from my button-clicking tasks. As I explained in Chapter 1, my reasons for using these kinds of tasks were primarily practical: they can be more easily administered at a distance, and they generate more easily analysable data. However, I will also say that the style of experiment I have favoured is one that relies, in general, on *simulating* the pressures that are involved in real language production, rather than allowing them to emerge organically. I do still fundamentally believe that we can learn something interesting from this approach, and indeed, that it affords us something important: the ability to pinpoint the precise mechanisms we are interested in, and control for other potential confounds. All experiments are models of the systems they are intended to study: they are designed to resemble the real systems in important ways, but abstract away from many of their vast complexities. Nonetheless, it is always worth considering how much abstraction is too much.

To bolster the conclusions presented in this thesis, it will be important to test how well my findings generalise to different types of production task. Although I can only speculate at this stage, I would expect the results of Chapters 2 and 3 to be stronger with a more naturalistic production task, and the results of Chapter 4 to be weaker. These predictions are based on my interpretation of the mechanisms involved in the different processes, and the timescale on which these can have noticeable effects. We know that some kinds of regularisation can happen very quickly, in as little as one generation (Singleton & Newport 2004), and individuals clearly acquire a very complex mental representation of their language(s) within the first few years of their lives. This suggests to me that the aspects of production that give rise to these effects are sufficiently challenging to take hold quickly, and the tasks I used were probably a conservative estimate of these challenges. Large-scale sound change, on the other hand,

takes many generations to accumulate, and I think this probably comes down to the fact that motor articulation is not challenging *enough* to have strong effects on a short timespan. To get round this issue, my simulation of the pressures involved in articulation was obviously designed to exaggerate production difficulty, in a way that is more difficult to achieve with an oral production task — at least, I certainly found it difficult to come up with stimuli that would be so troublesome to pronounce as to create a strong preference for the alternatives (Appendix 4.B).

5.2.2 Different participant populations

All the experiments presented in this thesis tested adult participants. All bar one tested L1 English speakers, and the one exception (Experiment 2 in Chapter 3, with L1 German speakers) still didn't go beyond the Indo-European family. I am interested in "universal" properties of human cognition, yet claims of universality may not be warranted on the basis of such a restricted sample (Blasi et al. 2022).

I will say that, for the phenomena I studied in Chapters 2 and 4 (which only used English-speaking participants), I wouldn't expect different speaker populations to behave dramatically differently. For some phenomena, there is a clear problem with relying on English speakers, in that their language experience might bias them in the direction predicted by our hypothesis in a way that experience with other languages would not. For example, English speakers do not provide a strong test of a hypothesised bias in favour of word order harmony, since English itself is a harmonic language; converging evidence from speakers of non-harmonic languages helps locate this bias in general principles of cognition rather than L1 transfer (Culbertson & Newport 2015; Culbertson et al. 2012, 2020). However, for my purposes, English is not a special case: all languages have systematic rules and regularities, and all languages have some words that sound more alike and some that are more distinctive. In Chapter 3, it clearly *was* important to check whether the null result in the first experiment was simply because case marking was too inaccessible for English speakers, but we ruled out this explanation when we obtained the same pattern of results from a population whose native language *does* have case.

I think the idea of extending these projects to children has much greater potential

to affect the results. For Chapter 2, we know that children tend to regularise more than adults (e.g. Austin et al. 2022; Hudson Kam & Newport 2005, 2009), although arguably, the memory load manipulation I used may be seen as making adults more child-like (Perfors 2012). For Chapter 3, children tend to be more exploratory than adults (Liquin & Gopnik 2022; Sumner et al. 2019), so we might expect them to entertain multiple hypotheses about the language's grammar more readily than adults. And for Chapter 4, children might be less concerned than adults about minimising effort (Tal et al. 2023), which could lead to an even weaker tendency toward the expected frequency trade-off. However, although evidence from child learners would clearly be an interesting addition to these projects, it is worth considering whether children are really the primary agents of language change; it's a big question that I can't possibly answer here, but it's certainly the subject of healthy debate in linguistics and cultural evolution more generally (e.g. Bybee 2010; Cournane 2017, 2019; Ferman & Karni 2010; Hudson Kam & Newport 2005; Lew-Levy & Amir 2024; Newport 2020; Raviv & Arnon 2018; Tal et al. 2023).

5.2.3 Alternative model architectures

Finally, although I'm a big fan of the exemplar models I've used in this thesis, other model architectures could help expand our understanding of the complexity of factors that shape language structure. I mentioned Bayesian models in Chapter 1: these would provide a good way to learn more about the interaction between effects that arise during online processing, and prior biases that constrain what can be acquired in the first place. Both exemplar and Bayesian models provide their agents with some kind of initial input, which inevitably places some constraints on the eventual outcomes; models of emergent communication between neural networks allow for much greater freedom in the form of the languages that develop (e.g. Chaabouni et al. 2019; Conklin & Smith 2023; Guo et al. 2022; Resnick et al. 2020). But the extension I would be most keen to see — particularly for Chapter 4 where the model is explicitly communicative — is a network theory[3] (e.g. Fagyal et al. 2010; Josserand et al. 2021; Ke et al. 2008; Lou-Magnuson & Onnis 2018). Network models incorporate many elements of

[3]Network theory is essentially just a more grounded (less abstract) version of graph theory.

real language communities which may meaningfully shape the transmission and selection of new linguistic variants, including community size, degree of connectedness between members of the community, and varying levels of influence associated with different individuals. Of course, this is not mutually exclusive with the other possibilities I raised: agents with any kind of internal representation could be embedded in a communicative network.

5.3 Conclusion

Language is characterised by its enormous diversity. Yet in many ways, all languages are remarkably alike — and all humans remarkably alike in the way we learn and use them. My goal in this thesis was to understand more about where these commonalities come from. I started from the premise that, to address this question, we need to think more seriously about the way we create language in real-time. Thus far, the emphasis in language evolution research has, for whatever reason, tended to be on learning and comprehension. Yet I have argued that the challenges associated with *producing* language have fundamental implications for its structure. And, through a combination of empirical methods, I've shown that language adapts to a complex interplay of different pressures — pressures to be easily learnable, understandable, *and* producible. I hope that my work in this thesis will pave the way for a stronger focus on language production, and how it can shape language at every level: learning, use, and evolution.

Bibliography

Aben, B., Stapert, S., & Blokland, A. (2012). About the distinction between working memory and short-term memory. *Frontiers in Psychology*, *3*.

Acheson, D. J., & MacDonald, M. C. (2009). Twisting tongues and memories: Explorations of the relationship between language production and verbal working memory. *Journal of Memory and Language*, *60*(3), 329–350.

Aitchison, J. (1996). Small steps or large leaps? Undergeneralization and overgeneralization in creole acquisition. In H. Wekker (Ed.), *Creole languages and language acquisition* (pp. 9–32). Mouton De Gruyter.

Alexandrov, A. A., Boricheva, D. O., Pulvermüller, F., & Shtyrov, Y. (2011). Strength of word-specific neural memory traces assessed electrophysiologically. *PLOS ONE*, *6*(8), e22999.

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990.

Ambrose, S. A., Bridges, M. B., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works*. Jossey-Bass.

Ann, J. (1996). On the relation between ease of articulation and frequency of occurrence of handshapes in two sign languages. *Lingua*, *98*(1), 19–41.

Archangeli, D., & Pulleybank, D. (1994). *Grounded phonology*. MIT Press.

Arnon, I. (2015). What can frequency effects tell us about the building blocks and mechanisms of language learning? *Journal of Child Language*, *42*(2), 274–277.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth – children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*(2), 107–129.

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. *Scientific Reports*, *14*(1), 5255.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, *18*(3), 249–277.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Version 2).

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208.

Baddeley, A., & Hitch, G. (1974). Working Memory. *Psychology of Learning and Motivation — Advances in Research and Theory*, *8*(100), 47–89.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454.

Baroni, A. (2014). On the importance of being noticed: The role of acoustic salience in phonotactics (and casual speech). *Language Sciences*, *46*, 18–36.

Barrett, J. A. (2006). Numerical simulations of the Lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bates, E., & MacWhinney, B. (1981). Second-language acquisition from a functionalist perspective: Pragmatic, semantic, and perceptual strategies. *Annals of the New York Academy of Sciences*, *379*(1), 190–214.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1–26.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bentz, C., & Winter, B. (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, *3*(1), 1–27.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, *80*(2), 290–311.

Bickerton, D. (1981). *Roots of language*. Language Science Press.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823.

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

BNC Consortium. (2007). British National Corpus, XML edition.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47.

Bock, J. K. (1995). Sentence production: From mind to mouth. *Speech, Language, and Communication*, 181–216.

Bock, J. K., Finlay, B., Freyd, J., Irwin, D., Keil, F., Kroch, A., Stemberger, J., Zacks, R., Billman, D., Mckinney, J., Ostrin, R., & Saffran, E. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, J. K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). Academic Press.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*(1), 47–67.

Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., & Gillam, R. B. (2010). What you hear and what you say: Language performance in Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism*, *13*(3), 325–344.

Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990a). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, *29*(5), 591–610.

Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., & Fisch, S. M. (1990b). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, *29*(5), 591–610.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the less-is-more hypothesis. *Language Learning*, *69*, 13–41.

Brown, A. S. (2012). *The tip of the tongue state*. Taylor & Francis.

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2021). Semantic cues in language learning: An artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50.

Buchsbaum, B. R., & D'Esposito, M. (2019). A sensorimotor view of verbal working memory. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior*, *112*, 134–148.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. John Benjamins.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford University Press.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. *Annual Meeting of the Berkeley Linguistics Society*, *23*(1), 378–388.

Carroll, R., Svare, R., & Salmons, J. C. (2013). Quantifying the evolutionary dynamics of German verbs. *Journal of Historical Linguistics*, *2*(2), 153–172.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333.

Cathcart, C. A. (2024). Multiple evolutionary pressures shape identical consonant avoidance in the world's languages. *Proceedings of the National Academy of Sciences*, *121*(27), e2316677121.

Cavalli-Sforza, L. L., & Feldman, M. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2019). Anti-efficient encoding in emergent communication. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32 (NeurIPS 2019)* (p. 11).

Chambers, J. K., & Schilling, N. (2018). *The handbook of language variation and change*. John Wiley & Sons.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934–1949.

Chater, N., & Christiansen, M. H. (2010). Language acquisition meets language evolution. *Cognitive Science*, *34*(7), 1131–1157.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430.

Chevrot, J.-P., Dugua, C., & Fayol, M. (2008). Liaison acquisition, word segmentation and construction in French: A usage-based account. *Journal of Child Language*, *36*(3), 557–596.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*(3), 542–551.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509.

Christiansen, M. H., & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*(2), 105–135.

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, *60*(1), 21–43.

Clark, E. V. (1988). On the logic of contrast. *Journal of Child Language*, *15*(2), 317–335.

Cleary, A. M. (2017). Tip-of-the-tongue states. *The Curated Reference Collection in Neuroscience and Biobehavioral Psychology*, 433–449.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology 1: Between the grammar and physics of speech* (pp. 283–333, Vol. 1). Cambridge University Press.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology. Human Perception and Performance*, *16*(3), 551–563.

Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, *89*(3), 183–213.

Conklin, H., & Smith, K. (2023). Compositionality with variation reliably emerges in neural networks. *The Eleventh International Conference on Learning Representations*.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, *2*(4), 589–602.

Cournane, A. (2017). In defense of the child innovator. *Micro-change and Macro-change in Diachronic Syntax*, *23*, 10.

Cournane, A. (2019). A developmental view on incrementation in language change. *Theoretical Linguistics*, *45*(3), 127–150.

Coussé, E., & Mengden, F. (Eds.). (2014). *Usage-based approaches to language change*. John Benjamins Publishing Company.

Cowan, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition*, *21*(2), 162–167.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, & S. Belleville (Eds.), *Progress in brain research* (pp. 323–338, Vol. 169). Elsevier.

Crain, S., Koring, L., & Thornton, R. (2017). Language acquisition from a biolinguistic perspective. *Neuroscience & Biobehavioral Reviews*, *81*, 120–149.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Education.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Croft, W. (2003). *Typology and universals*. Cambridge University Press.

Croot, K., Au, C., & Harper, A. (2010). Prosodic structure and tongue twister errors. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 433–460). De Gruyter Mouton.

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*(5), 310–329.

Culbertson, J. (2023). Artificial language learning. In J. Sprouse (Ed.), *Oxford handbook of experimental syntax* (pp. 271–300). Oxford University Press.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Culbertson, J., Franck, J., Braquet, G., Barrera Navarro, M., & Arnon, I. (2020). A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, *204*, 104392.

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268–293.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology*, *6*, 1964.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive biases, linguistic universals, and constraint-based grammar learning. *Topics in Cognitive Science*, *5*(3), 392–424.

Culbertson, J., & Wilson, C. (2013). Artificial grammar learning of shape-based noun classification. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Cuskley, C., Castellano, C., Colaiori, F., Loreto, V., Pugliese, M., & Tria, F. (2017). The regularity game: Investigating linguistic rule dynamics in a population of interacting agents. *Cognition*, *159*, 25–32.

Cuskley, C., & Kirby, S. (2013). Synesthesia, cross-modality, and language evolution. In *The Oxford handbook of synesthesia* (pp. 869–899). Oxford University Press.

Cuskley, C., Pugliese, M., Castellano, C., Colaiori, F., Loreto, V., & Tria, F. (2014). Internal and external dynamics in language: Evidence from verb regularity in a historical corpus of english. *PLOS ONE*, *9*(8), e102882.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. T. (2017a). Words cluster phonetically beyond phonotactic regularities. *Cognition*, *163*, 128–145.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017b). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8), 2149–2169.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86.

Davis, E., & Smith, K. (2023). The learnability and emergence of dependency structures in an artificial language. *Journal of Language Evolution*, *8*(1), 64–89.

Deese, J. (1984). *Thought into speech: The psychology of a language*. Prentice-Hall.

DeGraff, M. (1999). *Language creation and language change: Creolization, diachrony, and development*. MIT Press.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*(S1), 1–25.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Donegan, P., & Stampe, D. (2009). Hypotheses of natural phonology. *Poznań Studies in Contemporary Linguistics*, *45*(1), 1–31.

Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. *Child Development*, *92*(2), 609–625.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology (2006)*, *71*(4), 808–816.

Dziubalska-Kołaczyk, K. (2014). Explaining phonotactics using NAD. *Language Sciences*, *46*, 6–17.

Ellis, N. C. (2022). Second language learning of morphology. *Journal of the European Second Language Association*, *6*(1), 34–59.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Engelhardt, P. E., Corley, M., Nigg, J. T., & Ferreira, F. (2010). The role of inhibition in the production of disfluencies. *Memory & Cognition*, *38*(5), 617–628.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*, 37–64.

Everett, C. (2018). The similar rates of occurrence of consonants across the world's languages: A quantitative analysis of phonetically transcribed word lists. *Language Sciences*, *69*, 125–135.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, *120*(8), 2061–2079.

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, *37*(2), 378–394.

Fedzechkina, M., Chu, B., & Florian Jaeger, T. (2018). Human information processing shapes language change. *Psychological Science*, *29*(1), 72–82.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2011). Functional biases in language learning: Evidence from word order and case-marking interaction. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Miniature artificial language learning as a complement to typological data. In L. Ortega, A. E. Tyler, H. I. Park, & M. Uno (Eds.), *The usage-based study of language learning and multilingualism* (pp. 211–232). Georgetown University Press.

Fedzechkina, M., & Roberts, G. (2020). Learners sacrifice robust communication as a result of a social bias. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *42*.

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Ferman, S., & Karni, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. *PLOS ONE*, *5*(10), e13648.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*(1), 57–84.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. In B. H. Ross (Ed.), *Psychology of learning and motivation* (pp. 209–246, Vol. 49). Academic Press.

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and cognitive processes*, *21*(7), 1011–1029.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, *37*(8), 1565–1578.

Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review*, *19*(5), 921–928.

Finley, S., & Badecker, W. (2007). Towards a substantively biased theory of learning. *Annual Meeting of the Berkeley Linguistics Society*, 142–153.

Finley, S., & Badecker, W. (2008). Analytic biases for vowel harmony languages. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project.

Finley, S., & Badecker, W. (2010). Linguistic and non-linguistic influences on learning biases for vowel harmony. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Flego, S. (2022). *The emergence of vowel quality mutation in Germanic and Dinka-Nuer: Modeling the role of information-theoretic factors using agent-based simulation* [Doctoral dissertation, Indiana University].

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, D. Steriade, & R. Kirchner (Eds.), *Phonetically based phonology* (pp. 232–276). Cambridge University Press.

Foulkes, P. (1997). Historical laboratory phonology—investigating /p/>/f/>/h/ changes. *Language and Speech*, *40*(3), 249–276.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*(2), 151–167.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Franke, M., & Jäger, G. (2012). Bidirectional optimization from reasoning and learning in games. *Journal of Logic, Language and Information*, *21*(1), 117–139.

Franken, M. K., Acheson, D. J., McQueen, J. M., Eisner, F., & Hagoort, P. (2017). Individual variability as a window on production-perception interactions in speech motor control. *The Journal of the Acoustical Society of America*, *142*(4), 2007.

Frauenfelder, U. H., Baayen, R. H., & Hellwig, F. M. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, *32*(6), 781–804.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179–228.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, *90*(1), 58–89.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Gardner, R. A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology*, *70*(2), 174–185.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 94–126). MIT Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013a). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013b). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310.

Givón, T. (1979). *On understanding grammar*. Academic Press.

Givón, T. (1985). Function, structure and language acquisition. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 1005–1028). Lawrence Erlbaum Associates Inc.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four ... or is it two? *Memory*, *12*(6), 732–747.

Goldberg, A. (2005). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Goldberg, A., & Casenhiser, D. (2008). Construction learning and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 197–215). Routledge.

Goldberg, A., & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, *26*(4), 300–311.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, *28*(5), 501–518.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. *The proceedings of the 24th Annual Child Language Research Forum*, 124–138.

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition*, *107*(3), 1155–1164.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219–260.

Gómez, M. J. L., Molina, T. B., Benítez, P. P., & Santiago de Torres, J. (2007). Predicting proficiency in signed language interpreting: A preliminary study. *Interpreting*, *9*(1), 71–93.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186.

Gonzalez-Gomez, N., Poltrock, S., & Nazzi, T. (2013). A "bat" is easier to learn than a "tab": Effects of relative phonotactic frequency on infant word learning. *PLOS ONE*, *8*(3), e59601.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*(5), 425–430.

Greenberg, J. H. (1965). Some generalizations concerning initial and final consonant clusters. *Linguistics*, *3*(18), 5–34.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 73–113). MIT Press.

Grey, S., Williams, J. N., & Rebuschat, P. (2014). Incidental exposure and L3 learning of morphosyntax. *Studies in Second Language Acquisition*, *36*(4), 611–645.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3503–3514.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*(2), 261–290.

Guo, S., Ren, Y., Mathewson, K., Kirby, S., Albrecht, S. V., & Smith, K. (2022). Expressivity of emergent language is a trade-off between contextual complexity and unpredictability. *arXiv*.

Haftka, B. (1996). Deutsch ist eine V/2-Sprache mit Verbendstellung und freier Wortfolge. In E. Lang & G. Zifonun (Eds.), *Deutsch - Typologisch* (pp. 121–141). De Gruyter.

Hahn, M., Degen, J., & Futrell, R. (2021). Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review*, *128*(4), 726.

Hahn, M., Mathew, R., & Degen, J. (2022). Morpheme ordering across languages reflects optimization for processing efficiency. *Open Mind*, *5*, 208–232.

Hahn, M., & Yang, X. (2022). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences*, *119*(24), e2122604119.

Haider, H. (2020). VO-/OV-Base Ordering. In M. T. Putnam & B. R. Page (Eds.), *The Cambridge Handbook of Germanic Linguistics* (pp. 339–364). Cambridge University Press.

Hall, K. C., Hume, E., Jaeger, T. F., & Wedel, A. (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard*, *4*.

Hallam, M., Jordan, F. M., Kirby, S., & Smith, K. (2025). Predictive structure emerges during generalisation of kin terms to new referents. *OSF Preprints*.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44.

Haspelmath, M. (2008). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, *19*(1), 1–33.

Havron, N., & Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, *48*(2), 244–260.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *39*(6), 1041–1070.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, *44*(1), 45–75.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Heeschen, C. (1993). Morphosyntactic characteristics of spoken language. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C.-W. Wallesch (Eds.), *Linguistic disorders and pathologies: An international handbook* (pp. 16–34). De Gruyter Mouton.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*(3), 297–306.

Hockett, C. F. (1960). The origin of speech. *Scientific American*.

Holmberg, A. (2015). Verb Second. In T. Kiss & A. Alexiadou (Eds.), *Volume 1* (pp. 342–383). De Gruyter Mouton.

Holmes, V. M., & Dejean De La Bâtie, B. (1999). Assignment of grammatical gender by native speakers and foreign learners of french. *Applied Psycholinguistics*, *20*(4), 479–506.

Holtz, A., Kirby, S., & Culbertson, J. (2022). The influence of category-specific and system-wide preferences on cross-linguistic word order patterns. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Holtz, A., Kirby, S., & Culbertson, J. (2023). With or without a system: How category-specific and system-wide cognitive biases shape word order. *OSF Preprints*.

Hopman, E. W. M. (2022). *Modality matters: Generalization in second language learning after production versus comprehension practice* [Doctoral dissertation, University of Wisconsin-Madison].

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, *29*(6), 961–971.

Hopper, P. (1987). Emergent grammar. *Annual Meeting of the Berkeley Linguistics Society*, 139–157.

Horst, J., & Hout, M. (2014). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Unpublished manuscript*.

Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Language*, *91*, 906–937.

Hudson Kam, C. L. (2019). Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Language Learning and Development*, *15*(4), 317–337.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(3), 815–821.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, *59*(1), 30–66.

Hurford, J. R. (1999). The evolution of language and languages. In R. Dunbar, C. Knight, & C. Power (Eds.), *Evolution of culture*. Edinburgh University Press.

Hyman, R., & Jenkin, N. S. (1956). Involvement and set as determinants of behavioral stereotypy. *Psychological Reports*, *2*(3), 131–146.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, *24*(4), 541–577.

Jackendoff, R. (2002). Foundations of language. *Foundations of Language*.

Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, *1146*(1), 2–22.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335.

Jäger, G., & Rosenbach, A. (2008). Priming and unidirectional language change. *Theoretical Linguistics*, *34*(2), 85–113.

Jamet, D. (2009). A morphophonological approach to clipping in English. *Lexis*.

Jee, H., Tamariz, M., & Shillcock, R. (2022). Exploring meaning-sound systematicity in Korean. *Journal of East Asian Linguistics*, *31*(1), 45–71.

Johnson, J. S., Shenkman, K. D., Newport, E. L., & Medin, D. L. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language*, *35*(3), 335–352.

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. *Cognitive Science*, *44*(1), e12812.

Jordens, P., de Bot, K., & Trapman, H. (1989). Linguistic aspects of regression in German case marking. *Studies in Second Language Acquisition*, *11*, 179–204.

Josserand, M., Allassonnière-Tang, M., Pellegrino, F., & Dediu, D. (2021). Interindividual variation refuses to go away: A Bayesian computer model of language change in communicative networks. *Frontiers in Psychology*, *12*.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*(3), 402–420.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 149.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kamps, C., Ferdinand, V., & Kirby, S. (2014). The origins of regularity in language: Why coordination matters. In E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (Eds.), *The Evolution of Language: Proceedings of the 10th international conference* (pp. 457–458).

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528–558.

Kanwal, J. (2018). *Word length and the principle of least effort: Language as an evolving, efficient code for information transfer* [Doctoral dissertation, University of Edinburgh].

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General*, *126*, 278–287.

Karmiloff-Smith, A. (1981). *A functional approach to child language: A study of determiners and reference*. Cambridge University Press.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Ke, J., Gong, T., & Wang, W. S.-Y. (2008). Language change and social networks. *Communications in Computational Physics*, *3*(4), 935–949.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kenanidis, P., Dąbrowska, E., Llompart, M., & Pili-Moss, D. (2023). Can adults learn L2 grammar after prolonged exposure under incidental conditions? *PLOS ONE*, *18*(7), e0288989.

Keogh, A., Kirby, S., & Culbertson, J. (2024). Predictability and variation in language are differentially affected by learning and production. *Cognitive Science*, *48*(4), e13435.

Keogh, A., & Lupyan, G. (2024). Who benefits from redundancy in learning noun class systems? *Proceedings of the 15th International Conference on the Evolution of Language*.

Keppenne, V., Hopman, E. W. M., & Jackson, C. N. (2021). Production-based training benefits the comprehension and production of grammatical gender in L2 German. *Applied Psycholinguistics*, *42*(4), 907–936.

King, A., & Wedel, A. (2020). Greater early disambiguating information for less-probable words: The lexicon is shaped by incremental processing. *Open Mind*, *4*, 1–12.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford University Press.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review 2016 24:1*, *24*(1), 118–137.

Kirby, S., Cornish, H., & 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*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Klein, E. (1971). *A comprehensive etymological dictionary of the english language: Dealing with the origin of words and their sense development thus illustrating the history of civilization and culture*. Elsevier Publishing Company.

Koranda, M., Bulgarelli, F., Weiss, D. J., & MacDonald, M. C. (2020). Is language production planning emergent from action planning? A preliminary investigation. *Frontiers in Psychology*, *11*.

Koranda, M., Zettersten, M., & MacDonald, M. C. (2018). Word frequency can affect what you choose to say. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*.

Krevitt, B., & Griffith, B. C. (1972). A comparison of several Zipf-type distributions in their goodness of fit to language data. *Journal of the American Society for Information Science*, *23*(3), 220–221.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*(2), 119–131.

Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (p. 127). John Benjamins Publishing Company.

Lee, C., Lew-Williams, C., & Goldberg, A. (2022). Accessibility factors that lead to good-enough language production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Leeuw, J. R. d., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223–232.

Levitt, A. G., & Healy, A. F. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language*, *24*(6), 717–733.

Levshina, N. (2020). Efficient trade-offs as explanations in functional linguistics: some problems and an alternative proposal. *Revista da ABRALIN*, *19*(3), 50–78.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, 234.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447.

Lew-Levy, S., & Amir, D. (2024). Children as agents of cultural adaptation. *Behavioral and Brain Sciences*.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, *48*(4), 839–862.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Springer Netherlands.

Liquin, E. G., & Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. *Cognition*, *218*, 104940.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Lorch, M. P., & Meara, P. (1989). How people listen to languages they don't know. *Language Sciences*, *11*(4), 343–353.

Lou-Magnuson, M., & Onnis, L. (2018). Social network limits language complexity. *Cognitive Science*, *42*(8), 2790–2817.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Lüdecke, D. (2018). ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, *3*(26), 772.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLOS ONE*, *5*(1), e8559.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, *25*(1), 47–53.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35.

Macklin-Cordes, J. L., & Round, E. R. (2020). Re-evaluating phoneme frequencies. *Frontiers in Psychology*, *11*.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*(1), 199–227.

MacWhinney, B. (2018). A unified model of first and second language learning. In M. Hickmann, E. Veneziano, & H. Jisa (Eds.), *Sources of variation in first language acquisition: Languages, contexts, and learners* (pp. 287–312). John Benjamins Publishing Company.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.

Mahmoud, H. (2008). *Polya urn models*. CRC Press.

Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive Science*, *42*(8), 3116–3134.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Mahowald, K., James, A., Futrell, R., & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, *91*, 5–27.

Majerus, S. (2013). Language repetition and short-term memory: An integrative framework. *Frontiers in Human Neuroscience*, *7*.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–178.

Marks, E. A., Bond, Z., & Stockmal, V. (2003). Language experience and the representation of phonology in an unknown language. *Revista Espanola De Linguistica Aplicada*.

Marquet, P. A., Allen, A. P., Brown, J. H., Dunne, J. A., Enquist, B. J., Gillooly, J. F., Gowaty, P. A., Green, J. L., Harte, J., Hubbell, S. P., O'Dwyer, J., Okie, J. G., Ostling, A., Ritchie, M., Storch, D., & West, G. B. (2014). On theory in ecology. *BioScience*, *64*(8), 701–710.

Martin, A., Adger, D., Abels, K., Kanampiu, P., & Culbertson, J. (2024). A universal cognitive bias in word order: Evidence from speakers whose language goes against it. *Psychological Science*, *35*(3), 304–311.

Martin, A., Holtz, A., Abels, K., Adger, D., & Culbertson, J. (2020). Experimental evidence for the influence of structure and meaning on linear order in the noun phrase. *Glossa: a journal of general linguistics*, *5*(1).

Martin, A., & Peperkamp, S. (2020). Phonetically natural rules benefit from a learning bias: A re-examination of vowel harmony and disharmony. *Phonology*, *37*(1), 65–90.

Martin, A., & White, J. (2021). Vowel harmony and disharmony are not equivalent in learning. *Linguistic Inquiry*, *52*(1), 227–239.

Martin, R. C., & Romani, C. (1994). Verbal working memory and sentence comprehension: A multiple-components view. *Neuropsychology*, *8*(4), 506–523.

Martindale, C., Gusein-Zade, S. M., McKenzie, D., & Borodovsky, M. Y. (1996). Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics*, *3*(2), 106–112.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Meir, I., Aronoff, M., Börstell, C., Hwang, S.-O., Ilkbasaran, D., Kastner, I., Lepic, R., Ben-Basat, A. L., Padden, C., & Sandler, W. (2017). The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition*, *158*, 189–207.

Meisezahl, M., Kirby, S., & Culbertson, J. (2023). Variability and learning in language change: The case of v2.

Meylan, S. C., & Griffiths, T. L. (2024). Word forms reflect trade-offs between speaker effort and robust listener recognition. *Cognitive Science*, *48*(7), e13478.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Momma, S. (2021). Filling the gap in gap-filling: Long-distance dependency formation in sentence production. *Cognitive Psychology*, *129*, 101411.

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition*, *133*(3), 530–534.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, *16*(4), 495–500.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, Part II: Substance. *Language and Linguistics Compass*, *6*(11), 702–718.

Motamedi, Y., Wolters, L., Naegeli, D., Kirby, S., & Schouwstra, M. (2022). From improvisation to learning: How naturalness and systematicity shape language evolution. *Cognition*, *228*, 105206.

Motley, M. T., & Baars, B. J. (1975). Encoding sensitivites to phonological markedness and transitional probability: Evidence from spoonerisms. *Human Communication Research*, *1*(4), 353–361.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of speech, language, and hearing research: JSLHR*, *44*(4), 778–792.

Napoli, D. J., & Sutton-Spence, R. (2014). Order of the major constituents in sign languages: Implications for all language. *Frontiers in Psychology*, *5*, 376.

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., & Donkin, C. (2018). When extremists win: Cultural transmission via iterated learning when populations are heterogeneous. *Cognitive Science*, *42*(7), 2108–2149.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language. *Language Sciences*, *10*(1), 147–172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Newport, E. L. (2020). Children and adults as language learners: Rules, variation, and maturational change. *Topics in Cognitive Science*, *12*(1), 153–169.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *6*(312).

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In A. J. Wills & E. M. Pothos (Eds.), *Formal approaches in categorization* (pp. 18–39). Cambridge University Press.

Ohala, J. J. (1993). Coarticulation and phonology. *Language and Speech*, *36*(2), 155–170.

Pankratz, E. (2025). *Segmentation, rule formation, and the emergence of generalisation* [Doctoral dissertation, University of Edinburgh].

Papadopoulou, D., Varlokosta, S., Spyropoulous, V., Kaili, H., Prokou, S., & Revithiadou, A. (2011). Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research*, *27*, 173–205.

Parodi, T., Schwartz, B. D., & Clahsen, H. (2004). On the L2 acquisition of the morphosyntax of German nominals. *Linguistics*, *42*, 669–705.

Pavlov, P. I. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, *18*(3), 571–590.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(4), 486–506.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*(2), 138–155.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the emergence of linguistic structure* (pp. 137–157). John Benjamins Publishing Company.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Pitts Cochran, B., McDonald, J. L., & Parault, S. J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, *41*(1), 30–58.

Poizner, H., Newkirk, D., & Bellugi, U. (1983). Processes controlling human movement: Neuromotor constraints on American Sign Language. *Journal of Motor Behavior*, *15*(1), 2–18.

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38.

Pozdniakov, K., & Segerer, G. (2007). Similar place avoidance: A statistical universal. *Linguistic Typology*, *11*(2), 307.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in Cognitive Sciences*, *11*(7), 274–279.

Raviv, L., & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition*, *181*, 160–173.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863.

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition*, *206*, 104475.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In *The handbook of language emergence* (pp. 237–263). John Wiley & Sons, Ltd.

Resnick, C., Gupta, A., Foerster, J., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. *arXiv*.

Ribot, K. M., Hoff, E., & Burridge, A. (2018). Language use contributes to expressive language growth: Evidence from bilingual children. *Child Development*, *89*(3), 929–940.

Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1–30.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1), 107–142.

Rogers, M. (1987). Learners difficulties with grammatical gender in German as a foreign language. *Applied Linguistics*, *8*(1), 48–74.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.

Rohde, D. L. T., & Plaut, D. C. (2003). Less is less in language acquisition. In P. Quinlin (Ed.), *Connectionist modelling of cognitive development*. Psychology Press.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science*, *1*(6), 906–914.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*(1), 181–203.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition*, *35*(2), 261–290.

Saldana, C., Kirby, S., Truswell, R., & Smith, K. (2019). Compositional hierarchical structure evolves through cultural transmission: An experimental study. *Journal of Language Evolution*, *4*(2), 83–107.

Saldana, C., Smith, K., Kirby, S., & Culbertson, J. (2021). Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order. *Language Learning and Development*, *17*(2), 158–188.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: A comparison of online and mail survey modes. *Journal of Applied Econometrics*, *23*(3), 351–371.

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment*. Oxford University Press.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 3–32). Cambridge University Press.

Schouwstra, M., & De Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436.

Schwab, J. F., Lew-Williams, C., & Goldberg, A. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of child language*, *45*(5), 1054–1072.

Schwartz, B. L. (2002). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*. Lawrence Erlbaum Associates Publishers.

Schwering, S. C., & MacDonald, M. C. (2020). Verbal working memory as emergent from language comprehension and production. *Frontiers in Human Neuroscience*, *0*, 68.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, *12*(4), 323–328.

Senghas, A., Coppola, M., Newport, E. L., & Supalla, T. (1997). Argument structure in Nicaraguan Sign Language: The emergence of grammatical devices. In E. Hughes, M. Hughes, & A. Greenhill (Eds.), *Proceedings of the Boston University conference on language development* (pp. 550–561, Vol. 21). Cascadilla Press.

Seržant, I. A., & Moroz, G. (2022). Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved. *Humanities and Social Sciences Communications*, *9*(1), 1–9.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 41–55.

Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, *32*, 15–20.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–464.

Siegel, J. (2007). Recent evidence against the language bioprogram hypothesis. *Studies in Language*, *31*(1), 51–88.

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75.

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(3), 394–410.

Silverstein, M. (1976). Hierarchy of features and ergativity. In *Grammatical Categories in Australian Languages* (pp. 163–232). De Gruyter.

Sims-Williams, H. (2022). Token frequency as a determinant of morphological change. *Journal of Linguistics*, *58*(3), 571–607.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*(4), 370–407.

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Smith, A., Goffman, L., & Stark, R. E. (1995). Speech motor development. *Seminars in Speech and Language*, *16*(2), 87–98, quiz 98–99.

Smith, K. (2011). Learning bias, cultural evolution of language, and the biological evolution of the language faculty. *Human Biology*, *83*(2), 261–278.

Smith, K. (2020). How culture and biology interact to shape language and the language faculty. *Topics in Cognitive Science*, *12*(2), 690–712.

Smith, K., Ashton, C., & Sims-Williams, H. (2023). The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Smith, K., Bowerman, J., & Smith, A. D. M. (2024). Semantic extension in a novel communication system is facilitated by salient shared associations. *PsyArXiv Preprints*.

Smith, K., Brighton, H., & Kirby, S. (2003a). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems*, *06*(4), 537–558.

Smith, K., Kirby, S., & Brighton, H. (2003b). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371–386.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Smith, K. H. (1969). Learning co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 319–321.

Snyder, W. (2007). *Child language: The parametric approach*. Oxford University Press.

Spike, M. (2016). *Minimal requirements for the cultural evolution of language* [Doctoral dissertation, University of Edinburgh].

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2013). Learning, feedback and information in self-organizing communication systems. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 3442–3447.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive Science*, *41*(3), 623–658.

Stamp, R., Dachkovsky, S., Hel-Or, H., Cohn, D., & Sandler, W. (2024). A kinematic study of phonetic reduction in a young sign language. *Journal of Phonetics*, *104*, 101311.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, *8*(4), 339–356.

Steels, L. (2012). Self-organization and selection in cultural language evolution. In L. Steels (Ed.), *Experiments in cultural language evolution* (pp. 1–37). John Benjamins Publishing Company.

Steels, L., & Loetzsch, M. (2012). The grounded naming game. In L. Steels (Ed.), *Experiments in cultural language evolution* (pp. 41–59). John Benjamins Publishing Company.

Stemberger, J. P. (1990). Wordshape errors in language production. *Cognition*, *35*(2), 123–157.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1), 413–422.

Stockmal, V., Muljani, D., & Bond, Z. (1996). Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, *3*, 1748–1751 vol.3.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of speech, language, and hearing research : JSLHR*, *49*(6), 1175–1192.

Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211.

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32*(4), 827–853.

Sumner, E. S., Li, A. X., Perfors, A., Hayes, B. K., Navarro, D. J., & Sarnecka, B. W. (2019). The exploration advantage: Children's instinct to explore leads them to find information that adults miss. *PsyArXiv*, 11.

Swain, M. (2005). The output hypothesis: Theory and research. In *Handbook of Research in Second Language Teaching and Learning* (pp. 471–483). Routledge.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132.

Tal, S., Smith, K., Arnon, I., & Culbertson, J. (2023). Communicative efficiency is present in young children and becomes more adult-like with age. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Tamariz, M. (2008). Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, *3*(2), 259–278.

Tanaka, M. N., Branigan, H. P., McLean, J. F., & Pickering, M. J. (2011). Conceptual influences on word order and voice in sentence production: Evidence from Japanese. *Journal of Memory and Language*, *65*(3), 318–330.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, *105*(2), 300–333.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*(16), 4530–4535.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475.

Trott, S., & Bergen, B. (2022). Languages are efficient, but for whom? *Cognition*, *225*, 105094.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). *Irvine Phonotactic Online Dictionary* (Version 2.0).

van de Velde, H., Gerritsen, M., & van Hout, R. (1996). The devoicing of fricatives in standard Dutch: A real-time study based on radio recordings. *Language Variation and Change*, *8*(2), 149–175.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*(3), 285–316.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science*, *8*(4), 314–317.

Vigliocco, G., Vinson, D. P., Martin, R. C., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of Tip of the Tongue states and anomia. *Journal of Memory and Language*, *40*(4), 534–558.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of experimental psychology. Learning, memory, and cognition*, *28*(4), 735.

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: Effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514–529.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: Implications for spoken word recognition. *Frontiers in Psychology*, *2*.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S., & Luce, P. A. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, *52*(2), 193–204.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, *40*(1), 47–62.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1), 306–311.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & cognition*, *31*(4), 491–504.

Walsh, M., Möbius, B., Wade, T., & Schütze, H. (2010). Multilevel exemplar theory. *Cognitive Science*, *34*(4), 537–582.

Wasow, T., Perfors, A., & Beaver, D. I. (2005). The puzzle of ambiguity. In C. Orgun & P. Sells (Eds.), *Morphology and the web of grammar: Essays in memory of steven g. lapointe*. CSLI Publications.

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review*, *103*(4), 761–772.

Wedel, A. (2006). Exemplar models, evolution and language change. *The Linguistic Review*, *23*(3), 247–274.

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, *4*(4), 319–355.

Wedel, A., & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, *2*(1), 77–93.

Wells, J. C., & Hung, T. T. (1990). Longman pronunciation dictionary. *RELC Journal*, *21*(2), 95–97.

Wilshire, C. E. (1999). The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech*, *42*(1), 57–82.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960–967.

Winter, B., & Wedel, A. (2016). The co-evolution of speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in Cognitive Science*, *8*(2), 503–513.

Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology*, *58*(4), 221.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1–14.

Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. Clark-Cotton, & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development* (pp. 1–11).

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2006). Formulaic language. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics* (pp. 590–597). Elsevier.

Wu, S., Cotterell, R., & O'Donnell, T. (2019). Morphological irregularity correlates with frequency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5117–5126.

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Yadav, H., Mittal, S., & Husain, S. (2022). A reappraisal of dependency length minimization as a linguistic universal. *Open Mind*, *6*, 147–168.

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*(2), B45–B55.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, *81*, 103–119.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, *12*(1), 85–129.

Zheng, S., & Do, Y. (2025). Substantive bias in artificial phonology learning. *Language and Linguistics Compass*, *19*(1), e70005.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.