

Do memory limitations provide a selective pressure for languages to develop systematic structure? An experimental investigation

Exam number: B168648

Word count: 7,888

MSc Linguistics



THE UNIVERSITY
of EDINBURGH

School of Philosophy, Psychology and Language Sciences

August 2021

Abstract

Human languages universally rely on a system of compositionality, where complex utterances are decomposable into smaller pieces (words and morphemes) which correspond to different parts of the meaning of the whole. There is good evidence that languages in which this mapping between meanings and forms is more regular or predictable are easier to learn. But what is the link between regularity and learnability? Previous work has shown that working memory is central to language development, processing and use. We use an artificial language learning experiment to test whether memory limitations could also play a role in language evolution, by exerting a selective pressure on languages to develop more transparent morphological structures as they evolve. We find that more systematically structured languages are easier to learn, and all languages are harder to learn with reduced working memory capacity. However, we do not find any evidence that the advantage of structure is greater when memory is more limited, nor that the errors made by learners with less working memory capacity result in output languages that are more systematically structured than the input. We explore a range of possible explanations for these findings and suggestions for future work on the topic.

Keywords: working memory; language evolution; artificial language learning; morphological complexity

Contents

1	Introduction	3
1.1	Working memory	4
1.1.1	Developmental evidence	5
1.1.2	Clinical evidence	6
1.1.3	A role for working memory in language evolution?	6
1.2	The present study	7
2	Methods	7
2.1	Experiment design	7
2.2	Participants	7
2.3	Materials	8
2.4	Procedure	10
2.5	Analysis	11
2.5.1	Binary measure of accuracy	11
2.5.2	Similarity measure of accuracy	12
2.5.3	Language structure score	12
3	Results	12
3.1	Interference task	13
3.2	Main hypotheses	13
3.2.1	Overall learnability	13
3.2.2	Does structure provide a greater advantage under load?	15
3.2.3	Do learners generalise more under load?	16
4	Discussion	18
4.1	The non-linear effect of structure on learnability	18
4.2	The effect of structure does not depend on cognitive load	20
4.3	Cognitive load does not cause learners to eliminate irregularity	22
4.4	Final thoughts	23
	References	24
	Appendix	27

1 Introduction

Human languages are universally constructed compositionally: the meaning of an utterance is a function of the meaning of its parts and the relationship between them. Arguments have long been made in favour of a biological basis for this property of language (Hauser et al. 2002, Nowak et al. 2001, Pinker and Bloom 1990), but in fact, a large and growing body of work is demonstrating that cultural transmission may be all we need to explain the existence of such adaptively structured languages (Beckner et al. 2017, Kirby et al. 2008, Smith and Kirby 2008). Specifically, compositionality emerges as a trade-off between two competing pressures: learnability and expressivity (Kirby et al. 2015). The most easily learned languages are those in which every meaning is expressed by the same signal, what Kirby et al. term *degenerate* languages. However, these languages are functionally useless for communication since the intended meaning cannot be distinguished from all other possible meanings. The most communicatively expressive languages are those where there is no possibility of ambiguity because each meaning is associated with a unique form. Compositionality is one way of achieving this, but languages made up of random, idiosyncratic meaning-signal pairings in which complex meanings are not decomposable into smaller meaning-bearing units (*holistic* languages, Wray 1998) are equally expressive. So what prevents us from being able to learn and transmit such languages?

Of the two maximally expressive language types, holistic languages are considerably less compressible: any encoding of such a language would have to contain every possible complex meaning paired with a completely unique signal. For example, although the meanings “two big dogs” and “two small dogs” overlap significantly, there would be no systematic or predictable overlap in the linguistic forms used to communicate them in a holistic language. An encoding of a compositional language, on the other hand, would contain one entry for each constituent part of such meanings (“two”, “big”, “small”, “dog”, “-s”), with two implications: (1) simple meanings are not encoded multiple times for every compound meaning they appear in, and (2) these components can be recombined to communicate other similar meanings e.g. “two big bears”.

Between these two options, it seems self-evident that such an unwieldy system as the holistic language would be dispreferred, and indeed, we have good evidence that structured languages are easier to learn. In artificial language learning experiments, Kirby et al. (2008) showed that an increase in structure across generations correlated with a decrease in transmission error, and even in a single generation Raviv et al. (2021) found that highly structured languages were more faithfully reproduced. Even within the umbrella of compositionality, real human languages vary greatly in how they map meanings to morphosyntactic structures and the complexity of these mappings (Dryer & Haspelmath 2013), and there is evidence that more morphologically ‘transparent’ languages are more easily learned by children (Dressler 2003) and adult second language learners (DeKeyser 2005). But it is worth interrogating further what drives this link between compressibility and learnability.

One practical constraint is the cultural bottleneck on language transmission (e.g. Hurford 2000, Hurford 2002, Kirby 2002, Kirby 2017): in the real world, languages are learned from limited data, so those containing regularities which allow learners to generalise to unseen items are more likely to persist unchanged from one generation to the next. For example, if a language has only one way to mark pluralisation, then once a speaker of that language knows what a *fep* is, they automatically know how to talk about several of them. But more generally, it is widely assumed (e.g. Chater and Vitányi 2003, Hawkins 2004) that humans are naturally biased towards simplicity, and that more compressible languages are easier to learn under this principle. One hypothesis is that this is simply a universal prior

bias. But might we be able to identify specific cognitive mechanisms that could underpin the preference for simplicity in languages?

1.1 Working memory

Human learning and processing mechanisms undoubtedly constrain the ways in which languages evolve: language is shaped by the brain (Christiansen & Chater 2008). One such mechanism is working memory, the component of short-term memory used for temporary storage and processing of information in active attention. According to the original model (Baddeley & Hitch 1974), the working memory consists of a **central executive**, which allocates and retrieves information from other components, and two slave systems: the **phonological loop** which is specialised for processing verbally-coded information, and the **visuospatial sketchpad** which is specialised for processing visually-coded information. The phonological loop provides temporary storage of auditory information and prevents it from decaying too rapidly through a subvocal rehearsal process. Baddeley (2000) added a fourth component to the model: an **episodic buffer**, which integrates information from the two slave systems and long-term memory and passes it to the central executive (Figure 1).

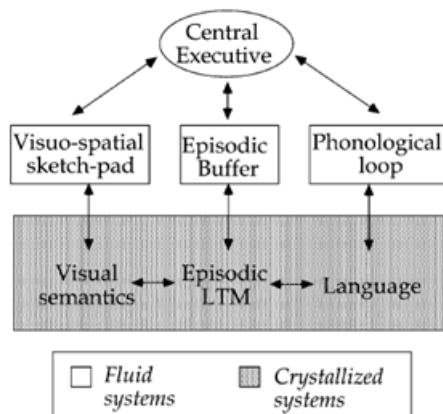


Figure 1: The revised working memory model (Baddeley 2003).

The Parallel Architecture perspective on language processing (Jackendoff 2007), on the other hand, considers linguistic working memory to consist of three subdivisions, one for each component of grammar (phonology, syntax and semantics), which operate simultaneously and pass information between themselves to decode the structure and meaning of the utterance to be processed. In setting out this model (Figure 2), Jackendoff argues that Baddeley's conception of the phonological loop, while sufficient to describe the processes behind memorisation of nonsense syllables, is inadequate for characterising the perception and processing of natural language.

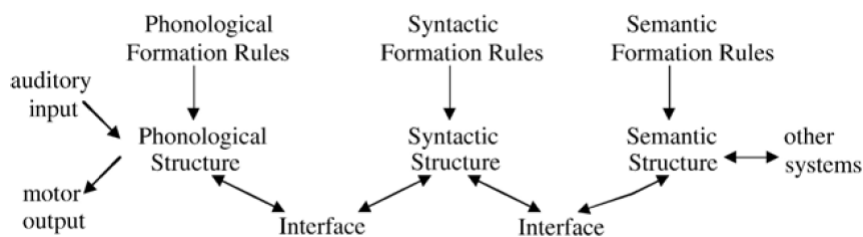


Figure 2: The Parallel Architecture (Jackendoff 2007)

Whatever our conception of the nature of working memory, a substantial body of work underlines the central role of such a system in human linguistic abilities.

1.1.1 Developmental evidence

The idea that there could be a causal relationship between working memory capacity and language learning has been widely explored in the developmental literature. The Less is More hypothesis (Newport 1988, 1990) argues that children are superior language learners *because of* (rather than in spite of) their relative cognitive deficits, as their limited memory capacity forces them to concentrate on smaller pieces of language which tend to be meaning-carrying and productive (words and morphemes).

This hypothesis has received considerable support from computational and experimental work. For example, Goldowsky and Newport (1993) construct a simulation in which a learner is tasked with identifying the morphemes of a language by counting form-meaning co-occurrences. The memory limitations of child learners are modelled as an input filter which randomly deletes half of the features making up each word in the initial stages, and gradually matures as the simulation progresses to preserve more and more of the input data. This simulation showed that, by forcing the model to focus on smaller units, the signal-to-noise ratio improves and the input filter actually optimises learning of regular items; exception words, on the other hand, are only learned once the filter becomes less restrictive. Elman (1993) had similar results in training recurrent neural networks to process complex sentences: the networks were *never* able to learn the language’s grammar when they came to the task with ‘adult-like’ capacities. Instead, training was only successful when the networks began with limited working memory and gradually matured.

Meanwhile, in the experimental domain, Cochran et al. (1999) found that, when teaching adults a novel sign language, limiting their working memory capacity impaired their vocabulary acquisition in the short-term, but prevented them from inferring the morphological system too quickly, yielding a long-term advantage in their ability to learn the language’s internal structure and generalise to new contexts. Kersten and Earles (2001) also found that adults were better able to learn the meanings and morphology of a miniature artificial language when they were forced to process small segments first before progressing to more complex stimuli. The importance of ‘starting small’ has also been demonstrated for adults learning recursion, both in artificial grammars (Lai & Poletiek 2010) and in foreign natural languages (Chin & Kersten 2010).

However, working memory limitations cannot account for all aspects of language development. Kam and Newport (2005) showed that, when trained on inconsistent input (e.g. a grammatical marker which appears only 70% of the time), adults will tend to probability match the observed inconsistency, while children will tend to regularise by producing the feature of interest either all the time or none of the

time. Perfors (2012) tested whether this effect could be explained by children’s lower working memory capacity but found that, while limiting adults’ memory capacity did reduce the number of nouns they were able to learn, it did not increase their tendency to regularise inconsistencies in the way these nouns were associated with determiners.

1.1.2 Clinical evidence

The relationship between working memory and language is also well-researched in the clinical domain. Evidence from aphasia patients shows that a preserved working memory system is crucial for language processing and production (see Ardila 2012, Wright and Shisler Marshall 2005 for reviews). For example, Haverkort (2005) argues that patients with Broca’s-type aphasia (characterised by laboured, agrammatic speech) retain their knowledge of grammar but select simpler syntax and shorter phrases because their production places less of a burden on their impaired working memory. Saffran and Marin (1975) report on a patient with conduction aphasia (characterised by load-sensitive speech repetition difficulties) who was unable to create mental representations of more complex constructions (such as passive or centre-embedded sentences) that would allow for even semantically consistent paraphrasing, much less verbatim repeating.

Language impairments are also common in other neurogenic conditions involving memory deficits (see Murray et al. 2001 for a review). For example, Moran and Gillon (2004) report that adolescents who suffered a traumatic brain injury in childhood performed poorly compared with their age-matched peers on language comprehension tasks, especially those which placed higher demands on working memory by requiring a large number of items to be retained and compared in real time. Almor et al. (1999) report that Alzheimer patients were less able than their age-matched peers to maintain in working memory the information necessary to process pronouns, and were consequently less sensitive to gender and number mismatches between pronouns and their referents.

1.1.3 A role for working memory in language evolution?

Christiansen and Chater (2016a, 2016b) argue that the fleeting nature of linguistic material in memory – what they term the “Now-or-Never” bottleneck – has fundamental implications for the nature of language and language change, including on an evolutionary timescale. Indeed, one of the features they attribute to this memory bottleneck is the presence of rule-like patterns in language, like the quasi-regularity of English past tense morphology where even so-called irregulars often follow identifiable sub-patterns (e.g. *sing* → *sang*, *ring* → *rang*, *spring* → *sprang*). The suggestion is that the more easily a particular chunk is processed in individual uses, the more entrenched it becomes as a systematic pattern, and the more likely it is to be passed down across generations of language learners and users.

Theories have also been posited on the co-evolution of language and working memory. For example, Coolidge and Wynn (2005, 2006, 2007) and Martín-Loeches (2006) argue for a relatively recent genetic mutation resulting in increased working memory capacity, which they believe would have significantly enhanced humans’ linguistic capabilities. And in a short comment paper, Coolidge (2012) proposed that grammatical structure is selected for as a means of bypassing the limits of working memory capacity as language evolves. However, as intuitively appealing as this claim is, it has never been substantiated experimentally, and this line of research has generally received scant attention in evolutionary linguistics.

1.2 The present study

The goal of the present study, therefore, is to test whether memory limitations could indeed exert a selective pressure on languages to become more structured as they evolve. This theory has two key implications:

1. Structure should confer a significant learning advantage when working memory capacity is reduced, over and above the advantage it provides as standard.
2. If the language to be learned is not already fully structured, it should become more so when working memory capacity is reduced i.e. learners should have an increased tendency to generalise where possible rather than preserving irregularity.

We use an artificial language learning experiment to test these hypotheses. For context, we also test overall learnability to confirm that structured languages are indeed easier to learn and that working memory limitations are (at least in the short-term) disadvantageous to word learning.

2 Methods

2.1 Experiment design

We use a 2x3 fully-crossed design to test the effects of working memory capacity and language structure on word learning. Working memory capacity is manipulated through an interference task designed to impose additional cognitive load on participants as they are learning words, reducing their ability to consolidate material through subvocal rehearsal. This interference task is modelled after the LOW CONCURRENT LOAD condition in Perfors (2012), albeit using digits rather than letters.

We use two levels of cognitive load:

- NO LOAD: Participants perform the word learning task only.
- LOAD: Participants perform a digit sequence recall task interspersed with the word learning task.

We use three levels of structure in participants' input language:

- STRUCTURED: The language's morphology is perfectly transparent: all affixes follow the same pattern.
- PARTIALLY STRUCTURED: The language's morphology is mostly transparent but with some exceptions: some affixes follow a pattern but some are unpredictable.
- UNSTRUCTURED: The language's morphology is opaque: affixes are completely unpredictable.

2.2 Participants

Participants were 160 self-reported adult native English speakers with no known memory or language disorders, recruited via Prolific. They were provided with a downloadable information sheet and gave informed consent to participate. The experiment lasted for up to 20 minutes ($M = 12.71$, $SD = 5.26$),

for which participants were paid £2.50 (average hourly pay = £12.45). Eight participants were excluded for the following reasons: self-reporting the use of written notes in an exit questionnaire contrary to the instructions (3), data saving errors (3), and failing to provide usable data on any critical trials (2)¹. We therefore analysed data from 152 participants, split by condition as shown in Table 1.

	No load	Load	Total
Structured	26	25	51
Partially structured	25	26	51
Unstructured	27	23	50
Total	78	74	152

Table 1: Number of participants per condition submitted to analysis.

2.3 Materials

Participants were asked to learn a small artificial language consisting of trisyllabic text labels paired with nine simple images. The images varied along two dimensions of shape and fill pattern (Figure 3).

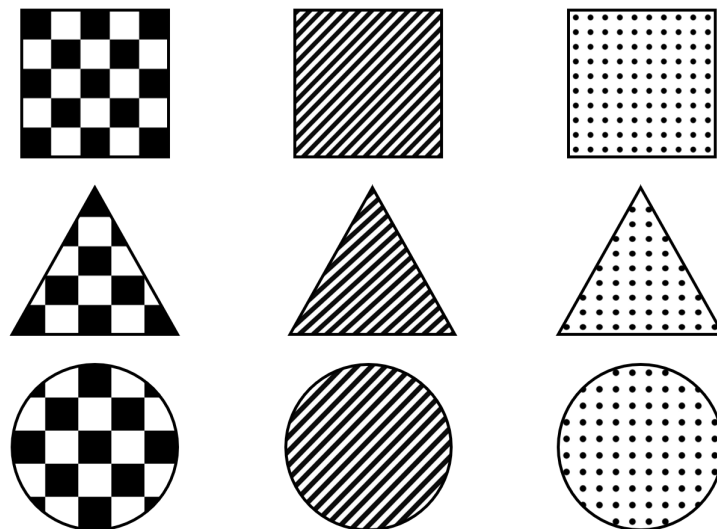


Figure 3: The full meaning space. Images varied along two dimensions: shape (square, triangle or circle) and fill pattern (checkerboard, striped or spotted). Although potentially more salient than fill pattern, colour was avoided for accessibility reasons. Feedback was gathered during piloting to ensure that the images were still sufficiently discernible.

The training language for each participant was generated according to their condition, by concatenating CV syllables chosen from two unique sets of nine: {“ga”, “be”, “tu”, “pu”, “ki”, “hu”, “mo”, “la”, “ne”} and {“ni”, “do”, “lu”, “va”, “su”, “bi”, “fa”, “ri”, “pa”}. All 81 possible combinations were generated in advance and manually checked by the experimenter for English words².

Each word consisted of a bisyllabic stem for the shape (formed by reduplication of the syllables in

¹Either by only typing digits rather than letters, or by typing the same string on every trial.

²Only unambiguously English words were excluded; potential homophones (e.g. “puni” ~ *punny*) were allowed on the grounds that they could not be identified objectively when considering different accents and possible pronunciations of the artificial words. Therefore, the only combination that was disallowed was “lava”.

the first set) and a monosyllabic suffix for the fill pattern (chosen from the second set). To make the languages more learnable, shapes were labelled consistently in all conditions; the only element that varied by condition was the systematicity of the suffixes³. Thus, all languages were generated by first randomly selecting and duplicating one syllable per shape from the first set. In UNSTRUCTURED conditions, each syllable from the second set was then randomly assigned to one of the nine objects and paired with the relevant stem. In STRUCTURED conditions, one syllable per fill pattern was randomly chosen from the second set and combined with the stems systematically. In PARTIALLY STRUCTURED conditions, first a structured language was generated as above, and then two positions in the meaning space were randomly chosen as exceptions such that one shape and one fill pattern in the meaning space followed a consistent pattern, and the other two of each were associated with two systematically structured words and one exception. Exception words were generated by randomly pairing two of the remaining syllables in the second set with the relevant stems and assigning them to the selected positions. Example languages are given in Table 2.













	Structured			Partially Structured			Unstructured		
									
	gaga-ni	gaga-do	gaga-lu	gaga-va	gaga-do	gaga-lu	gaga-va	gaga-do	gaga-ni
	bebe-ni	bebe-do	bebe-lu	bebe-ni	bebe-su	bebe-lu	bebe-su	bebe-bi	bebe-fa
	tutu-ni	tutu-do	tutu-lu	tutu-ni	tutu-do	tutu-lu	tutu-ri	tutu-pa	tutu-lu

Table 2: Example languages for each condition (hyphens are included here for clarity and not displayed to participants). In these languages, “gaga” means *square*, “bebe” means *triangle* and “tutu” means *circle*. In the structured language, “ni” means *checkerboard*, “do” means *striped* and “lu” means *spotted*; the two elements are then combined predictably to create structured words for the complex meanings. The partially structured language starts with the same structured language, but then exception words are assigned to the two shaded positions; note that, as there is never more than one exception word in a given row or column, there is enough information in the non-shaded cells for the structured language to be perfectly reconstructed. In the unstructured language, the stems are randomly combined with all nine suffix syllables such that there is no systematic structure on the fill dimension.

The meaning space was chosen to keep the language small enough to be at least partially learnable in all conditions, but large enough that a structured language still has an advantage over an unstructured one in terms of compressibility. We use as a measure of compressibility the ratio of meaning-bearing units in the language to items in the meaning space. Given a two-dimensional meaning space with n features along the first dimension and m along the second where the n features are labelled consistently and the m features vary by condition, the number of meaning-bearing units in the language is given by $n + m$ for a structured language, or $(n \times m) + n$ for an unstructured language. An unstructured language will therefore always have more meaning-bearing units than items in the meaning space, a ratio in this case of 12:9, whereas a structured language in the present meaning space has a ratio of 6:9; we assume that this smaller inventory of unique elements confers a learning advantage (Kirby et al. 2008).

³It is worth noting that all languages in this design could be described as compositional, since words are always decomposable into smaller, meaning-bearing units: even in UNSTRUCTURED conditions, there are still identifiable stems and the suffixes are still *meaningful*, just not predictably so. When we refer to ‘language structure’, therefore, we are referring primarily to the systematicity of the language’s morphology, rather than whether or not the language exhibits compositionality. In this sense, these languages correspond more closely to the kind of morphological paradigms we see in real languages, which vary widely in their degree of regularity but are never truly holistic.

In a 2x3 meaning space, on the other hand, the number of meaning-bearing units only differs by three depending on language type, so there is presumably less advantage in learning a structured language over an unstructured one.

2.4 Procedure

The experiment was written in JavaScript using the JsPsych library (de Leeuw 2015) and ran fullscreen in participants' web browser.

Participants were randomly assigned to one of the six conditions at the start of the experiment. Participants in all conditions were told that they would be taught a small part of a new language and then tested to see how much they had learned. Additional instructions in the LOAD conditions were modelled after Perfors (2012): participants in these conditions were told that we were interested in how well people can learn words when the task is difficult, so they would also be asked to memorise and recall short sequences of numbers while they were learning the words. They were told that they would be given feedback throughout on their performance on this task.

The experiment consisted of a training phase and a testing phase. All participants saw the full set of stimuli in both phases. In the training phase, participants were given six passes through the stimuli to learn the language, with the order of presentation randomised in each pass. In each training trial, participants observed as an image was presented for 1000ms and then its text label appeared for a further 4000ms.

In LOAD conditions, participants completed an additional sequence recall task sandwiched around the word learning task. Before being shown each image-label pair, a pseudo-random sequence of three digits was displayed for 2500ms and participants were asked to memorise 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 neighboured 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. Immediately after the image-label pair disappeared, participants were asked to retype the numbers they had just memorised, in order. As in Perfors (2012), participants were given feedback on the number of digits they had recalled in the correct position and how long they had taken to respond. A schematic of a training trial is given in Figure 4.

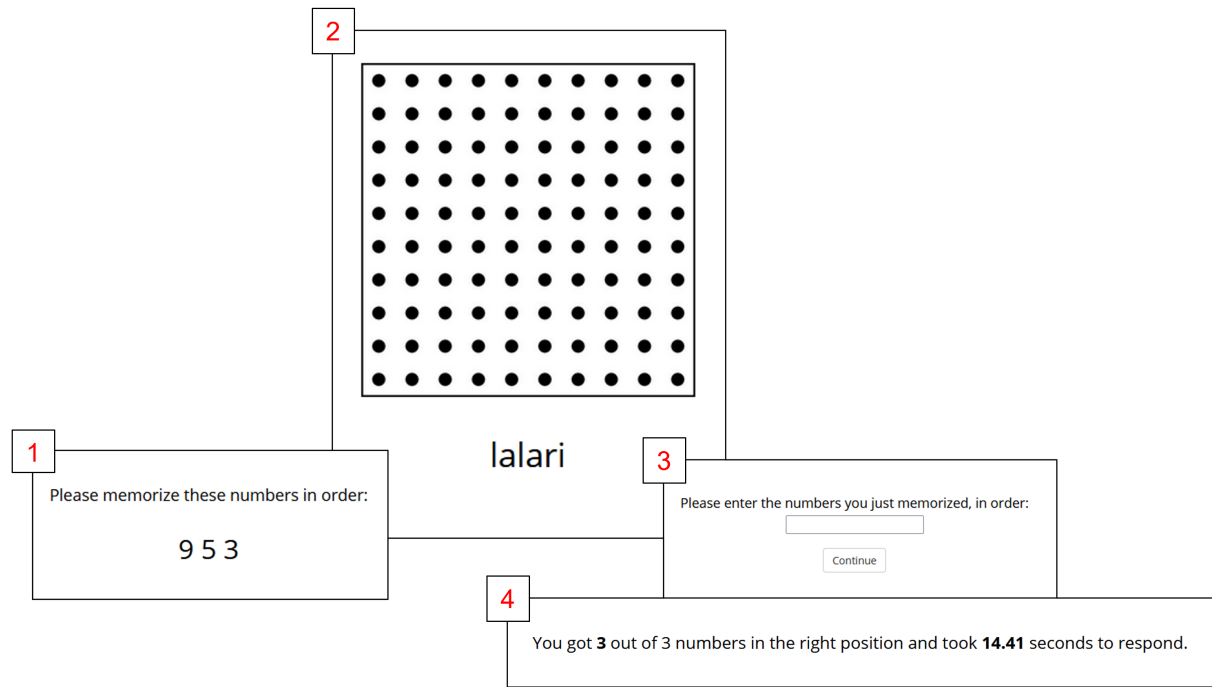


Figure 4: Participant view of a training trial. Participants in LOAD conditions saw all four of these elements on each trial. Participants in NO LOAD conditions saw only the second element.

In all conditions, participants were first given one practice round to make sure they understood the task. The stimulus for this trial was chosen randomly from the set of nine. In between each training pass, participants were given a break of 5000ms before the experiment continued.

In the testing phase, participants in all conditions were shown all nine images in a random order and given as long as they needed to type in their labels. No feedback was provided on accuracy.

2.5 Analysis

Participants in LOAD conditions were given a score out of three on each training trial for the number of digits they had recalled in the correct position. The following measures were additionally calculated for all participants.

2.5.1 Binary measure of accuracy

This measure simply indicates whether a participant's response on a given testing trial was correct or incorrect. If the label typed by the participant was an exact match for the target label (case insensitive, and after removing any spaces and punctuation), the trial was marked 1; otherwise it was marked 0.

2.5.2 Similarity measure of accuracy

This measure quantifies how closely a participant’s response on a given testing trial resembled the target label for that stimulus, and is calculated as follows:

$$\text{similarity}(p, t) = 1 - \frac{LD(p, t)}{M}$$

where $LD(p, t)$ is the Levenshtein distance (Levenshtein 1966) between the label produced by the participant p and the target label t (i.e. the minimum number of insertions, deletions or substitutions needed to transform one label into the other), normalized for length by dividing by M , the length of the longer of the two labels. Subtracting this number from 1 yields a measure of similarity where high scores correspond to more faithful reproductions of the input language (a score of 1 being a perfect match), while low scores indicate that the participant has produced a very different label from the one they learned.

2.5.3 Language structure score

This measure quantifies the degree of systematic structure in a particular language according to the procedure in Kirby et al. (2008). First, a matrix is constructed of the normalized Levenshtein distance between all pairs of strings in the language. A second matrix is constructed by calculating the distance between all pairs of meanings, so that meanings differing in one feature (shape or fill) have a distance of 1, and meanings differing in both features have a distance of 2. These two distance matrices are then submitted to a Mantel test (Mantel 1967) using the *culvevo* package (Stadler 2018) in R (R Core Team 2020). This test calculates the Pearson’s product-moment correlation between the two set of distances, giving an indication of the extent to which similar meanings are expressed by similar labels, and then compares it against a null distribution generated by performing the same calculation on 1,000 random permutations of the two matrices. The correlation coefficient of the observed mapping is then z -scored with respect to this sample to determine a 95% confidence interval above which a language is considered to associate labels and meanings in a non-random way; a language with a score of 1.96 or above is therefore taken to be significantly structured. Each participant received two scores on this measure: one for their input language, and one for the language they produced in testing. The measure is undefined for languages which use the same label for all meanings (or all but one), as all reorderings of such a language are equally structured.

3 Results

We analysed the data in R (R Core Team 2020), using the *lme4* package (Bates et al. 2015) to generate mixed effects models. Where p -values are reported for fixed effects as a whole, these were generated using likelihood ratio tests to compare the full model with a reduced model lacking that fixed effect. Where p -values are reported for the comparison between particular levels of a fixed effect, these were generated using the *lmerTest* package (Kuznetsova et al. 2017).

3.1 Interference task

Before proceeding to the main analysis, we need to first verify that the interference task works as expected; if participants in LOAD conditions opted to try and improve their performance on the language learning task by ignoring the digit sequences, comparison between conditions on this dimension would be rendered meaningless.

Fortunately, overall performance on the digit sequence recall task was close to ceiling ($M = 2.75$, $SD = 0.68$). We performed a mixed effects linear regression predicting the number of digits correct with a fixed effect of language structure and random intercepts for participant and target sequence. While inspection of the means by condition suggests that there were small differences (Table 3), language structure was not a significant predictor of performance ($\chi^2(2) = 1.499$, $p = 0.473$). We are therefore satisfied that participants in all conditions were attending to this task.

Language structure	M	SD
Structured	2.79	0.65
Partially structured	2.72	0.72
Unstructured	2.76	0.66

Table 3: Mean accuracy (digits recalled in the correct position, out of three) on the digit sequence recall task by structure condition. Performance was close to ceiling across the board with no significant difference between conditions ($p = 0.473$).

3.2 Main hypotheses

All full models in Sections 3.2.1 and 3.2.2 included fixed effects for cognitive load (dummy-coded, with NO LOAD as reference level) and language structure (Helmert-coded⁴, with STRUCTURED as reference level), as well as a random intercept for participant. Other models are specified in Section 3.2.3; where used, cognitive load and language structure followed the same coding scheme. All models were tested with an additional random intercept for stimulus but failed to converge, so this term was removed.

3.2.1 Overall learnability

We predicted that language learning would be made more difficult by cognitive load and lack of compositional structure in the input. Figure 5 shows performance by condition on both the binary and similarity measures of accuracy.

⁴Helmert coding compares the mean of each level of a categorical variable to the mean of subsequent levels of the variable. In this case, we compare STRUCTURED languages to the combination of PARTIALLY STRUCTURED and UNSTRUCTURED languages, and then compare PARTIALLY STRUCTURED and UNSTRUCTURED languages. This coding scheme allows us to determine whether languages with perfect systematicity are different from those with *any* irregularity, and then to test whether there is any difference between different degrees of irregularity. This was felt to be a more meaningful comparison than dummy coding, which would simply have compared PARTIALLY STRUCTURED and UNSTRUCTURED languages to STRUCTURED ones, and not to each other.

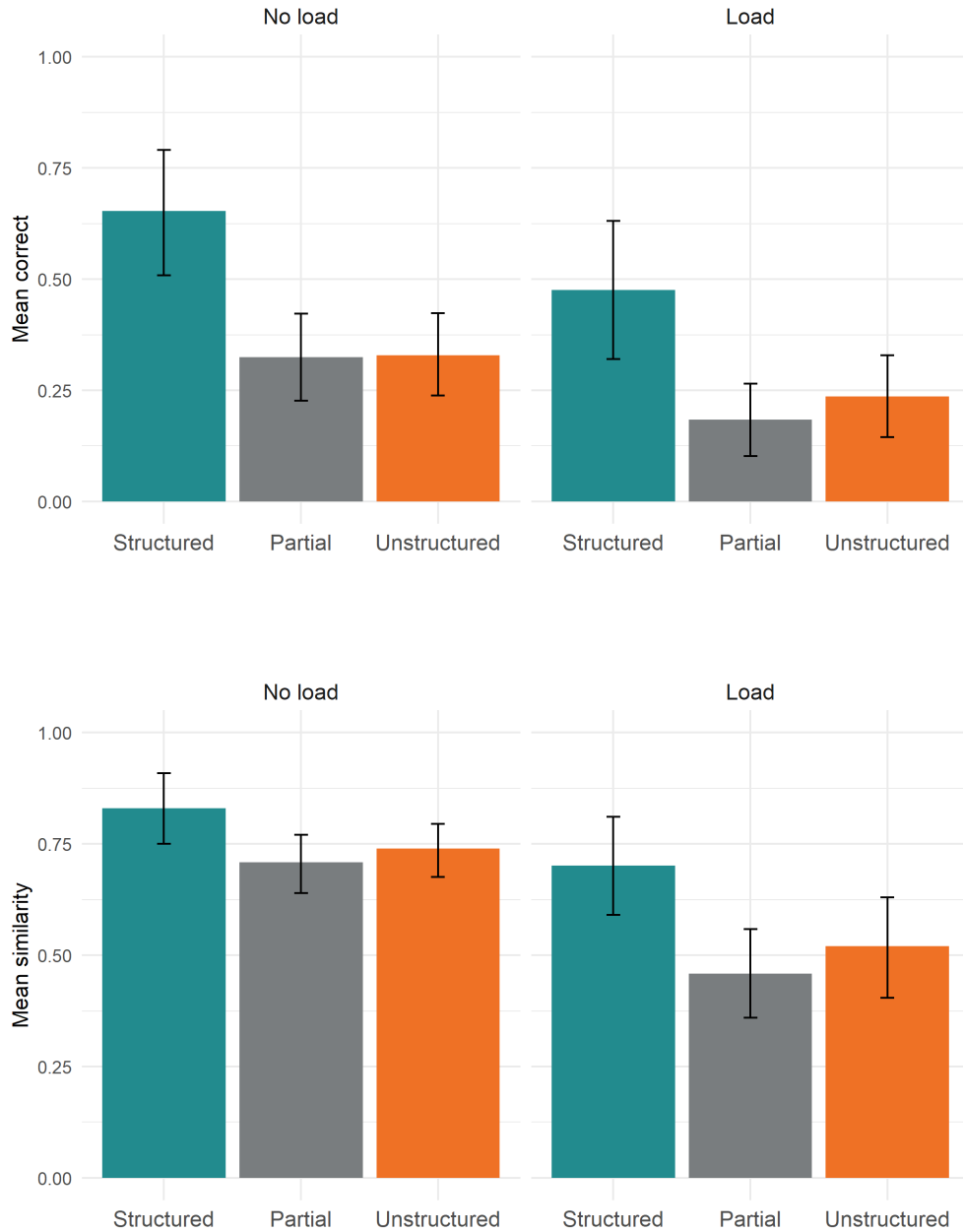


Figure 5: Accuracy on testing trials by condition (top: binary measure, bottom: similarity measure). Error bars represent bootstrapped 95% confidence intervals over the mean by participant. Unsurprisingly, performance is generally lower on the binary measure, which awards the same score (0) to a label that is only one letter different from the target and one that bears no resemblance to the target. Performance in NO LOAD conditions is higher on both measures than for the corresponding LOAD conditions. Performance on STRUCTURED languages is higher on both measures than for other languages, but the difference between PARTIALLY STRUCTURED and UNSTRUCTURED languages appears to be negligible.

As expected, performance on both measures is highest in the STRUCTURED/NO LOAD condition, and lower under cognitive load for all levels of language structure. However, the effect of language structure appears to be non-linear, with the lowest performance on both measures observed in the PARTIALLY STRUCTURED/LOAD condition; indeed, for both measures of accuracy and regardless of cognitive load,

performance is lower for PARTIALLY STRUCTURED languages than for UNSTRUCTURED ones. Nonetheless, it is the case that performance is higher across the board on STRUCTURED languages than on languages containing any amount of randomness, and the difference between PARTIALLY STRUCTURED and UNSTRUCTURED languages appears to be minimal.

On the binary measure of accuracy, a mixed effects logistic regression showed a significant effect of cognitive load, with decreased log odds of a correct response in LOAD conditions ($\beta = -1.058$, $SE = 0.347$, $z = -3.047$, $p < 0.01$). Model comparison also showed a significant effect of language structure ($\chi^2(2) = 33.419$, $p < 0.001$); inspection of the coefficients reveals that performance in STRUCTURED conditions was significantly better than in other conditions ($\beta = 2.128$, $SE = 0.378$, $z = 5.635$, $p < 0.001$), but with no significant difference between PARTIALLY STRUCTURED and UNSTRUCTURED conditions ($\beta = -0.275$, $SE = 0.420$, $z = -0.654$, $p = 0.513$).

On the similarity measure, a mixed effects linear regression showed a significant effect of cognitive load, with lower accuracy in LOAD conditions ($\beta = -0.199$, $SE = 0.038$, $t = -5.178$, $p < 0.01$). Model comparison also showed a significant effect of language structure ($\chi^2(2) = 15.372$, $p < 0.001$); inspection of the coefficients reveals that performance in STRUCTURED conditions was significantly better than in other conditions ($\beta = 0.158$, $SE = 0.041$, $t = 3.888$, $p < 0.001$), but with no significant difference between PARTIALLY STRUCTURED and UNSTRUCTURED conditions ($\beta = -0.048$, $SE = 0.047$, $t = -1.008$, $p = 0.315$).

Thus, in line with our predictions, cognitive load and language structure were both reliable predictors of performance on the language learning task.

3.2.2 Does structure provide a greater advantage under load?

We predicted that language structure would provide more of an advantage when working memory capacity was reduced. We should therefore expect to additionally find an interaction between cognitive load and language structure alongside their main effects, such that the differential between learning success in LOAD and NO LOAD conditions is greater for UNSTRUCTURED languages than for STRUCTURED ones. However, although it appears from Figure 6 that this may be the case on the similarity measure of accuracy, the interaction term was not significant on either this measure ($\chi^2(2) = 1.797$, $p = 0.407$) or the binary measure ($\chi^2(2) = 0.349$, $p = 0.840$). Thus, we have insufficient evidence to support our hypothesis that the effect of language structure on learning success depends on the amount of working memory capacity available to learners.

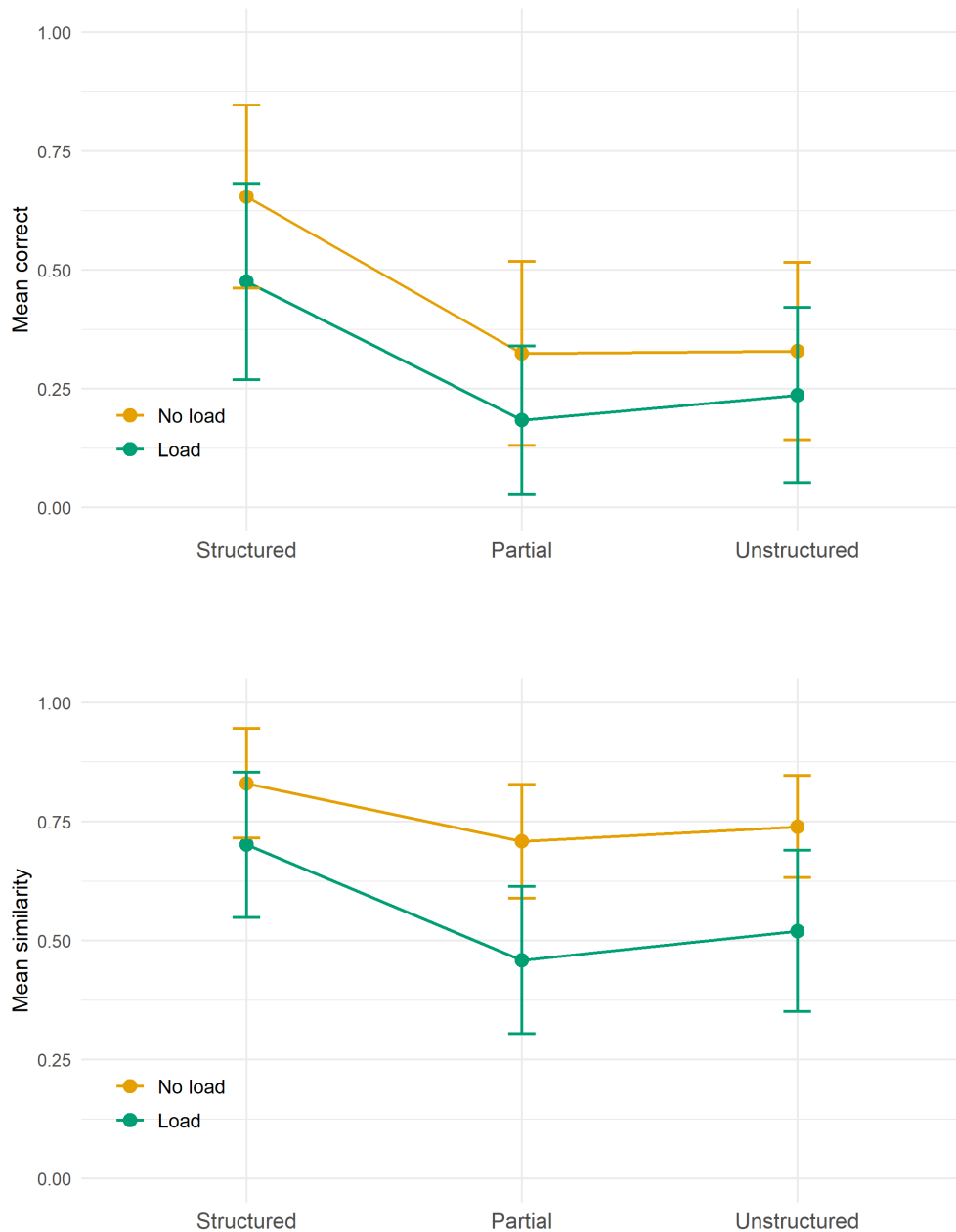


Figure 6: The differential between learning success in LOAD and NO LOAD conditions by language structure: binary measure (top) and similarity measure (bottom). Error bars represent 95% confidence intervals over the mean by trial. The effect does seem to be going in the predicted direction on the similarity measure, where there is a smaller difference between load conditions for STRUCTURED languages than others. However, we can see from the substantial overlap in error bars that there is too much noise to determine whether the effect of load differs by level of language structure: indeed, the interaction term is non-significant on both measures (binary measure: $p = 0.840$; similarity measure: $p = 0.407$).

3.2.3 Do learners generalise more under load?

We predicted that cognitive load would increase learners' tendency to generalise where possible rather than learning exception words holistically. We should therefore expect the output languages of LOAD

learners to score more highly on our structure measure than those of NO LOAD learners. We would also expect any increase in structure to be most notable in the PARTIALLY STRUCTURED/LOAD and UNSTRUCTURED/LOAD conditions.

Figure 7 shows that, in fact, structure decreases across the board, and more steeply for LOAD conditions, with only the most highly structured input languages remaining significantly structured in testing for these learners.

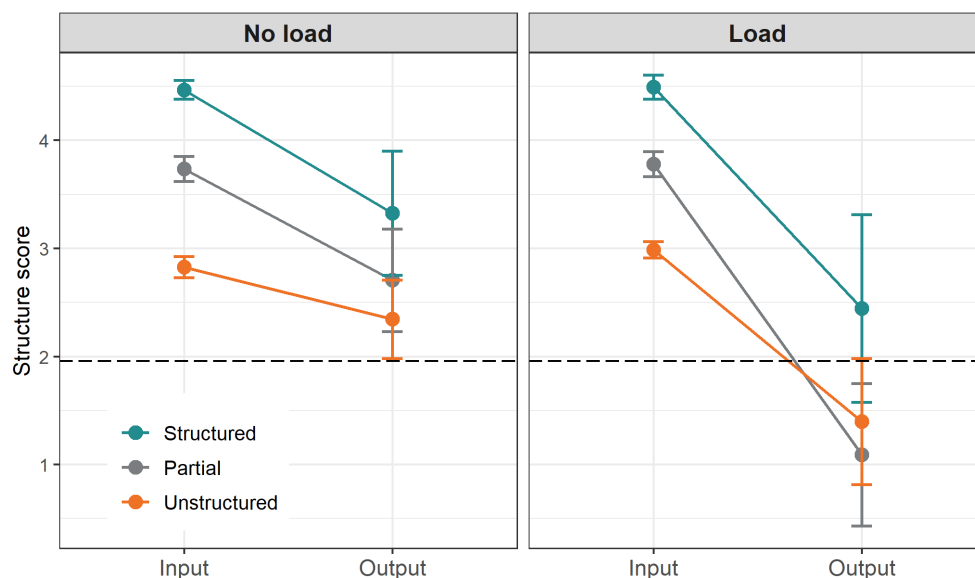


Figure 7: Structure scores of input and output languages by condition. Error bars represent 95% confidence intervals over the mean by participant. The dotted line ($z = 1.96$) shows the 95% confidence interval that the observed language could be obtained by random assignment of labels to meanings; languages above this threshold are therefore taken to be significantly structured. It is worth noting that all input languages in this setup are deemed to be structured, regardless of condition: this is because, as detailed in Section 2.3, all languages label shapes consistently, with the only variation between condition being in the regularity of the suffixes used to label fill patterns. The key thing is that the order is preserved and detectable by the structure measure: STRUCTURED languages score most highly, UNSTRUCTURED languages have the lowest scores and PARTIALLY STRUCTURED languages are in between.

A linear regression of structure scores by cognitive load, language structure and input/output (dummy-coded, with input as reference level) confirmed that output languages were significantly less structured than input languages ($\beta = -1.488$, $SE = 0.128$, $t = -11.664$, $p < 0.001$). A linear regression on the structure scores of output languages by cognitive load and language structure additionally confirmed that LOAD learners produced languages that were significantly less structured than those of NO LOAD learners ($\beta = 2.787$, $SE = 0.167$, $t = 16.682$, $p < 0.001$). This model also showed that output languages in STRUCTURED conditions were significantly more structured than those in other conditions ($\beta = 1.004$, $SE = 0.253$, $t = 3.961$, $p < 0.001$), but with no significant difference between PARTIALLY STRUCTURED and UNSTRUCTURED conditions ($\beta = 0.031$, $SE = 0.294$, $t = 0.104$, $p = 0.917$), despite the fact that these language types were different in the input. Thus, cognitive load is a reliable predictor of the degree of compositional structure in the output language, but in the opposite direction than predicted: in other words, the way LOAD learners are failing to learn their input languages is not by generalising. Furthermore, learning is not successful enough in any condition to preserve the structure of the input

languages.

Since these results were such a poor match for our predictions, we decided to additionally model the regularity of individual words in the input language as predictor of learning success by load condition. Our hypothesis relies on LOAD learners being unsuccessful specifically at learning exception words, and significantly worse than NO LOAD learners on these items, so if this is not the case then we should not be surprised that we have not found the effect we predicted. We coded each input word as either regular (0) if the suffix appeared more than once in that language, or irregular (1) if the suffix was unique in that language. We then constructed two models predicting accuracy with fixed effects of cognitive load, irregularity and the interaction between these terms, and a random intercept for participant. On the binary measure, a mixed effects logistic regression revealed that irregularity was a significant predictor of accuracy, with decreased log odds of a correct response for irregular items ($\beta = -1.105$, $SE = 0.337$, $z = -3.279$, $p < 0.01$). However, the interaction term was not significant ($\chi^2(1) = 0.285$, $p = 0.593$). On the similarity measure, a mixed effects linear regression showed that learning success was lower for irregular items, but this effect was not significant ($\beta = -0.041$, $SE = 0.034$, $t = -1.192$, $p = 0.234$). The interaction term was also not significant ($\chi^2(1) = 0.163$, $p = 0.686$). Thus, we have mixed evidence on whether participants were less successful at learning irregular items overall, but no evidence to suggest that the effect of irregularity on accuracy was conditioned by cognitive load.

4 Discussion

As we predicted, reduced working memory capacity and language irregularity were both disadvantageous to word learning. However, we were unable to find any evidence supporting our prediction that the learning advantage provided by structured languages would be greater under cognitive load. And on our prediction that memory limitations should cause languages to become more structured, the evidence supported the opposite conclusion: structure decreased more dramatically under cognitive load.

That we replicate the well-attested effects of cognitive load and language structure on word learning should give us some confidence in our experiment design. In particular, the fact that the interference task clearly disrupted learning suggests that it was at least not too *easy* (we address the question of whether it was too hard in Section 4.2). However, we are left with a number of unexpected findings to account for.

4.1 The non-linear effect of structure on learnability

If we consider each stem-shape pair and each suffix-fill pair a learner sees as one piece of information, participants are receiving, every round of training, 18 pieces of information about the structure of the language. In UNSTRUCTURED conditions, half of these data points fit a pattern (the stems) and half are random (the suffixes). In STRUCTURED conditions, all 18 fit the pattern. In PARTIALLY STRUCTURED conditions, 16 data points fit the pattern and only two are random: in other words, these languages are really very close to being perfectly systematic. It is therefore quite remarkable that, not only is there no significant difference between the learnability of UNSTRUCTURED and PARTIALLY STRUCTURED languages (on either accuracy measure, and regardless of cognitive load), but looking at what little difference does exist suggests that the relationship between language structure and learnability follows a U-shape, where the hardest languages to learn are those with a mix of regular and irregular items.

Raviv et al. (2021) found a similar pattern⁵ and we share their intuitions about why this might be. If we assume that learners are (explicitly or implicitly) testing hypotheses about the language’s structure as training progresses, the data they receive may change their learning strategy. Specifically, perhaps learners in STRUCTURED conditions spot that there is a pattern fairly early on and subsequently focus on learning the rules that generated their data. Learners in UNSTRUCTURED conditions, meanwhile, may also spot that the stems follow a pattern but realise that the suffixes are arbitrary, and so focus on rote learning these once they have established the rule that generated the stems. Learners in PARTIALLY STRUCTURED conditions, however, are getting conflicting information throughout. We can imagine that these learners initially think they are learning a regular language – indeed, as quantified above, there is a lot of evidence for this hypothesis – and adopt the rule-learning strategy. In this case, exception words are going to be very disruptive, forcing learners to try and form a new rule even though the rule they were previously testing worked for nearly all the data. Interestingly, we might expect this effect not to be so dramatic for LOAD learners, who presumably have limited capacity for explicit introspection about rules and exceptions.

As noted by Raviv et al., this finding is potentially problematic for iterated learning models of language evolution (Kirby et al. 2008, Kirby et al. 2015 etc.) which assume that structure gradually emerges over generations by virtue of the fact that *any* regularity is better than none. However, we suggest that our results are not at odds with this premise because our participants were not asked to generalise to unseen items, nor to use their language in a communication task: the advantage of structure (even partial) would no doubt be greater in these circumstances.

We should also consider the possibility that the reduction in learning success we have attributed to language structure could actually be explained by a confound: namely, the number of distinct morphemes in the input language (since a STRUCTURED language has only three unique suffixes, a PARTIALLY STRUCTURED language has five and an UNSTRUCTURED language has nine). Perhaps the low performance in PARTIALLY STRUCTURED conditions is therefore best explained by a combination of exception words disrupting a rule-learning strategy *and* the greater number of morphemes to be learned, and this effect is diluted in UNSTRUCTURED conditions precisely because these learners are focusing on memorisation rather than pattern-finding. To more conclusively test whether the regularity of a language’s morphology is the relevant factor in learnability, we could instead use an experiment design where the number of distinct suffixes is kept constant across conditions, but their relationship to meaning features is made more or less predictable (Table 4). However, this could (paradoxically) make the languages even harder to learn if participants think they have spotted another pattern (the same suffixes keep appearing) but cannot work out what conditions it.

⁵Although only on their binary measure of accuracy: the relationship between language structure and performance on their similarity measure was exponential rather than U-shaped. The discrepancy may be attributable to the larger amount of variation in the structure of input languages in their experiment (unstructured languages were truly random, rather than randomness being confined to a single morpheme of each word), but further testing would be required.

	Partially Structured			Unstructured		
	m1	m2	m3	m1	m2	m3
n1	x	y	z	x	y	z
n2	x	z	z	z	x	y
n3	x	y	x	y	z	x

Table 4: Given a two-dimensional meaning space with three features along the first dimension n and three along the second dimension m , languages with partial or complete irregularity could be constructed as seen here while keeping the number of suffixes constant between conditions. In the PARTIALLY STRUCTURED condition, the pattern is that suffix x corresponds to the first meaning on dimension m , suffix y corresponds to the second meaning and suffix z corresponds to the third meaning; the shaded cells are exceptions to this rule, but are still formed from the set of suffixes $\{x, y, z\}$. In the UNSTRUCTURED condition, each suffix from the set $\{x, y, z\}$ corresponds to a different meaning on dimension m depending on the meaning on dimension n ; thus, the association between form and meaning is completely unpredictable without adding more distinct morphemes to the language.

4.2 The effect of structure does not depend on cognitive load

We expected – but were unable to show – that language structure would provide a more significant learning advantage for learners with reduced working memory capacity than for those with full capacity.

There are many explanations we could consider for this null result. Most obviously, it is entirely possible that we find no effect because there really is no effect. Our hypothesis was motivated predominantly by the need to test a theoretical claim, not by previous experimental findings, so our lack of evidence may in fact be evidence that this claim is wrong and the cultural bottleneck presented in Section 1 is all we need to explain the emergence of linguistic structure. A slightly less drastic but related possibility is that working memory limitations are not a factor in the emergence of this type of structure, but may be in others. We find structure at all levels of language, and it is worth acknowledging that Coolidge was more interested in argument structure and word order than morphological complexity. Perhaps we have simply assumed his theory to apply more widely than intended. In any case, we should clearly discount all other possible explanations before abandoning this line of enquiry entirely.

Leaving aside the more existential considerations, there are also a number of factors in the experiment design that could have influenced our result. Firstly, we assumed that individual differences in participants’ working memory capacity would be accounted for by the random intercept in our models, and somewhat masked in the first place by random allocation of participants to conditions. However, the random intercept tells us little about the source of systematic differences between participants, and it is possible that we ended up with enough high performers in the ‘most difficult’ conditions (PARTIALLY STRUCTURED/LOAD and UNSTRUCTURED/LOAD) to wash out any effect that might have existed. It is certainly true that there was more variation in performance in LOAD conditions, suggesting that the interference task was more disruptive for some participants than others; this could be at least partly down to the amount of working memory they had available to start with. Future work could look to incorporate a measure of participants’ working memory capacity (e.g. a complex span task, Conway et al. 2005) as an additional predictor, bearing in mind the trade-off between collecting as much data as possible and keeping participants’ attention for long enough to ensure the quality of that data⁶.

⁶Indeed, the latter consideration was the main factor in our decision to omit such a task.

As foreshadowed above, another possibility is that the interference task was too hard, resulting in a floor effect in LOAD conditions so that not even a compositional language was learnable. It is difficult to take a view on this point for two reasons. Firstly, we did not have a specific expectation of how successful LOAD learners would be on this task, or how much *less* successful they would be than NO LOAD learners, because the cognitive load component is a new addition to the paradigm. Secondly, we have not attempted the experiment with a less demanding version of the interference task to examine the effect of varying the level of cognitive load on overall learnability. In fact, we assumed that our version of this interference task would be easier than the one it was modelled after (Perfors 2012) because the numeric domain should provide less of a conflict with the linguistic task than letter sequences, digits are sampled from a smaller pool than letters, and there is evidence that number sequences are better remembered than sequences of other verbal material because of the greater prevalence of random number sequences in natural language (phone numbers, dates/times etc., Jones and Macken 2015). In early piloting, we found little difference in performance on the word learning task between participants presented with three-digit sequences and those presented with six-digit sequences (indeed, there appears to be little difference between these levels in Perfors (2012) as well) so we assumed that the effect of cognitive load would lie primarily in the presence or absence of a task, rather than the relative difficulty of that task, but future work could explore this further, for example by testing the most minimal possible version of the interference task i.e. presenting only one digit to be recalled.

Conversely, it is plausible that the interference task was actually boosting performance in LOAD conditions by acting as an attention check. Performance on the digit sequence recall was high enough in all conditions that we know participants were paying attention to it, and therefore looking at their screens throughout the experiment. On the other hand, the training phase in NO LOAD conditions was entirely passive, so participants in these conditions may have lost concentration and consequently been less successful in learning the words than would otherwise have been the case. Again though, it is difficult to set a level of expected success to judge these participants by. And in any case, this line of reasoning would perhaps be more appropriate if we had failed to find a main effect of cognitive load, rather than as an explanation for the lack of interaction between cognitive load and language structure. Nonetheless, recent studies have shown that attention checks increase data quality and statistical power in online behavioural research (e.g. Aust et al. 2013, Oppenheimer et al. 2009), so may be worth considering for future work.

We should also consider what factors might predict performance on this word learning task other than concentration and the variables we have already tested. Most fundamentally, the task may simply be too hard in the time available to expect participants, on average, to ever exceed the performance threshold they have reached here. Although we took steps to make it easier (small meaning space, regular stems in all languages, multiple rounds of training), we could consider simplifying the design even further in a future experiment. Specifically, the fact that the testing phase was a free-typing exercise has created a lot of noise in the data: many of the labels produced by participants did not even match the input for number of characters, for example. It is possible, therefore, that there would have been an effect if we had more data or less variation in the data: after all, larger sample sizes are generally required to detect interaction effects (Leon & Heo 2009). A forced choice task – with participants asked to choose the label they were shown for a given stimulus in training from an array – would be one way to reduce noise and thus increase power to detect an effect if one exists. Anecdotally (and perhaps unsurprisingly), several participants commented in their debrief questionnaires that they would have preferred a multiple choice test. Alternatively, as an intermediate step, participants could be presented with the relevant stem for

each stimulus and asked to type only the suffix. We should also consider the possibility that our attempt to make things easier on participants by limiting the size of the meaning space may have inadvertently washed out any effect because the baseline advantage provided by a structured language is relatively small in this setup (as discussed in Section 2.3). However, it is difficult to know whether expanding the meaning space is the right way to go given that our hypotheses all rely on participants in NO LOAD conditions being able to learn their input language relatively well, and they were, on average, not even able to learn nine words successfully.

4.3 Cognitive load does not cause learners to eliminate irregularity

We also predicted that learners in LOAD conditions would be more likely to generalise, thus producing more highly structured output languages. However, contrary to this hypothesis, language structure actually decreased in all conditions, and significantly more so for LOAD learners.

In order to make sense of this finding, we need to consider what it would mean for structure to increase, given that all input languages in our setup were significantly structured to start with ($z > 1.96$). For a language's structure score to increase in UNSTRUCTURED conditions, participants would have had to get every stem right (or get them all wrong in the same, rule-governed way) and introduce some kind of pattern to the suffixes. And the criteria for structure to increase in PARTIALLY STRUCTURED conditions is even stricter: participants in these conditions would have had to get every stem and all seven regular suffixes right *and* make at least one of the irregular suffixes fit the pattern. Put this way, it is clear that we were expecting far too much of participants – especially in LOAD conditions where we predicted (correctly) that word learning would be less successful generally. The problem becomes particularly apparent when we look at the most highly structured output language from the PARTIALLY STRUCTURED/LOAD condition (Table 5): we would hardly expect anyone to do much better than this under cognitive load, but structure has still not increased. In fact, with input languages that scored so highly on our structure measure, any small mistakes would inevitably reduce structure – hence structure scores were not stable even in STRUCTURED conditions – and it is therefore obvious with hindsight that we could have anticipated this result.










	Input			Output		
						
	bebe-pa	bebe-bi	bebe-lu	baba-pu	baba-si	baba-su
	momo-su	momo-bi	momo-lu	momo-bu	momo-si	momo-su
	nene-su	nene-ri	nene-lu	nene-lu	nene-si	nene-lu

Table 5: The most highly structured output language from the PARTIALLY STRUCTURED/LOAD condition compared to the language the participant was asked to learn. Hyphens are added for clarity and were not displayed to or typed by the participant. Highlighted cells in the input denote the exception words. We can see that, in some ways, this participant *did* introduce more regularity to the language: all striped shapes have the same suffix in the output, and all checkerboard shapes at least end with the same letter. However, a handful of tiny inconsistencies mean that structure was essentially flat between input ($z = 3.741$) and output ($z = 3.738$).

We did initially attempt a version of this experiment where both shapes and fill patterns were labelled according to structure condition (i.e. words were genuinely holistic in UNSTRUCTURED conditions), assuming that the pressure for structure would be greater with this greater degree of irregularity; in this

setup, there would have been more scope for structure to increase since it would have started out much lower. However, performance was so low in piloting on even the easier conditions that we abandoned this design, judging that there was no possibility of structure increasing if participants found the task so hard that their responses were completely random.

One potential solution would be to add a generalisation element to the test, forcing participants to label unseen items. Indeed, this seems to be where the advantage of memory limitations in extrapolating rules comes into play in previous work (explored in Section 1.1.1), rather than in the initial memorisation of the training stimuli. This does introduce a confound: we already know that limiting data input will increase structure because of the bottleneck effect (Hurford 2000, Hurford 2002, Kirby 2002, Kirby 2017). Nonetheless, we could compare between conditions to test whether there is a more significant increase in structure under cognitive load, or in an iterated learning experiment (Kirby et al. 2008), whether output languages become significantly structured in fewer generations when memory is limited.

The forced choice task described above in Section 4.2 could be another way to test this hypothesis. Presented with an array of three labels – one correct, one incorrect but fitting the pattern of other items, and one random distractor – would LOAD learners be more likely to choose the regular item? This would allow us to capture any difference in tendency to eliminate irregularity much more simply and unambiguously than the Mantel test method. On the other hand, the relative ease of this task compared to a free-typing task could wash out any effect if learning success was near ceiling in all conditions, so it would be crucial to calibrate the size of the meaning space very precisely to ensure that the effects of cognitive load and language structure could still be detected.

4.4 Final thoughts

In summary, we have presented an experiment design where cognitive load and irregularity can both be shown to disrupt language learning. However, we have been unable to find an interaction between these two variables in support of our hypothesis that memory limitations could provide a selective pressure for the emergence of compositionality. Still, we have laid the groundwork for future investigation of the relationship between working memory and language structure, and offered a number of suggestions to advance this line of enquiry based on some of the shortcomings of the present study.

References

- Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., & Tyler, L. K. (1999). Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's Disease. *Brain and Language*, 67(3), 202–227.
- Ardila, A. (2012). Interaction between lexical and grammatical language systems in the brain. *Physics of Life Reviews*, 9(2), 198–214.
- Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior research methods*, 45, 527–535.
- Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11), 417–423.
- Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, 36(3), 189–208.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (47–89). Academic Press.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- 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.
- Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1), 19–22.
- Chin, S. L., & Kersten, A. W. (2010). The application of the less is more hypothesis in foreign language learning. *Proceedings of the 32nd annual conference of the cognitive science society* (150–155).
- 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, e62.
- Cochran, B. P., 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.
- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, 12(5), 769–786.
- Coolidge, F. L. (2012). On the emergence of grammatical language as a means of bypassing the limitations of working memory capacity: Comment on “Interaction between lexical and grammatical language systems in the brain” by Alfredo Ardila. *Physics of Life Reviews*, 9(2), 217–218.
- Coolidge, F. L., & Wynn, T. (2005). Working memory, its executive functions, and the emergence of modern thinking. *Cambridge Archaeological Journal*, 15, 5–27.
- Coolidge, F. L., & Wynn, T. (2007). Did a small but significant change in working memory capacity empower modern thinking? In P. Mellars, K. Boyle, C. Stringer, & O. Bar-Yosef (Eds.), *Rethinking the human revolution: New behavioural and biological perspectives on the origin and dispersal of modern humans* (79–90). Cambridge University Press.

- DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? a review of issues. *Language Learning*, 55(S1), 1–25.
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12.
- Dressler, W. (2003). Morphological typology and first language acquisition: Some mutual challenges. *Proceedings of the Fourth Mediterranean Morphology Meeting*, 4, 7–20.
- Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. <https://wals.info/>
- Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1), 71–99.
- Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The Less Is More Hypothesis. In E. V. Clark (Ed.), *The Proceedings of the 25th Annual Child Language Research Forum* (124–138). Cambridge University Press.
- Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.
- Havorkort, M. (2005). Linguistic representation and language use in aphasia. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (57–68). Routledge.
- Hawkins, J. (2004). *Efficiency and complexity in grammars*. Oxford University Press.
- Hurford, J. R. (2000). Social transmission favours linguistic generalisation. In C. Knight, M. Studdert-Kennedy, & J. Hurford (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (324–352). Cambridge University Press.
- Hurford, J. R. (2002). Expression/induction models of language evolution: Dimensions and issues. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (301–344). Cambridge University Press.
- Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, 1146, 2–22.
- Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, 144, 1–13.
- Kam, C. L. H., & 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.
- Kersten, A. W., & Earles, J. L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language*, 44(2), 250–273.
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (173–204). Cambridge University Press.
- Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin and Review*, 24, 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., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.

- Lai, J., & Poletiek, F. H. (2010). The impact of starting small on the learnability of recursion. *Proceedings of the 32nd annual conference of the cognitive science society* (1387–1392).
- Leon, A. C., & Heo, M. (2009). Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational Statistics and Data Analysis*, 53(3), 603–608.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10, 707.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2), 209–220.
- Martín-Loeches, M. (2006). On the uniqueness of humankind: Is language working memory the final piece that made us human? *Journal of Human Evolution*, 50(2), 226–229.
- Moran, C., & Gillon, G. (2004). Language and memory profiles of adolescents with traumatic brain injury. *Brain Injury*, 18, 273–88.
- Murray, L., Ramage, A., & Hopper, A. (2001). Memory impairments in adults with neurogenic communication disorders. *Seminars in speech and language*, 22, 127–36.
- 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.
- Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291(5501), 114–118.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872.
- Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, 67(4), 486–506.
- Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4), 707–727.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210, 104620.
- Saffran, E. M., & Marin, O. S. (1975). Immediate memory for word lists and sentences in a patient with deficient auditory short-term memory. *Brain and Language*, 2, 420–433.
- Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 3591–3603.
- Stadler, K. (2018). *cultevo: Tools, measures and statistical tests for cultural evolution* [R package version 1.0.2]. <https://kevinstadler.github.io/cultevo/>
- Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, 18(1), 47–67.
- Wright, H., & Shisler Marshall, R. (2005). Working memory in aphasia: Theory, measures, and clinical implications. *American Journal of Speech-Language Pathology*, 14, 107–18.
- Wynn, T., & Coolidge, F. L. (2006). The effect of enhanced working memory on language. *Journal of Human Evolution*, 50, 230–231.

Appendix

Supplementary information (including experiment code, stimuli, anonymised data and analysis code) can be downloaded from [**https://github.com/akkeogh/Dissertation**](https://github.com/akkeogh/Dissertation).