Behavior Research Methods (2021) 53:1218–1239
https://doi.org/10.3758/s13428-020-01473-6

Mini Pinyin: A modified miniature language for studying language learning and incremental sentence processing

Zachariah R. Cross[1] · Lena Zou-Williams[1] · Erica M. Wilkinson[1,2] · Matthias Schlesewsky[1] · Ina Bornkessel-Schlesewsky[1]

Published online: 6 October 2020

Abstract

Artificial grammar learning (AGL) paradigms are used extensively to characterise (neuro)cognitive bases of language learning. However, despite their effectiveness in characterising the capacity to learn complex structured sequences, AGL paradigms lack ecological validity and typically do not account for cross-linguistic differences in sentence comprehension. Here, we describe a new modified miniature language paradigm – Mini Pinyin – that mimics natural language as it is based on an existing language (Mandarin Chinese) and includes both structure and meaning. Mini Pinyin contains a number of cross-linguistic elements, including varying word orders and classifier-noun rules. To evaluate the effectiveness of Mini Pinyin, 76 (mean age = 24.9; 26 female) monolingual native English speakers completed a learning phase followed by a sentence acceptability judgement task. Generalised mixed effects modelling revealed that participants attained a moderate degree of accuracy on the judgement task, with performance scores ranging from 25% to 100% accuracy depending on the word order of the sentence. Further, sentences compatible with the canonical English word order were learned more efficiently than non-canonical word orders. We controlled for inter-individual differences in statistical learning ability, which accounted for ~20% of the variance in performance on the sentence judgement task. We provide stimuli and statistical analysis scripts as open-source resources and discuss how future research can utilise this paradigm to study the neurobiological basis of language learning. Mini Pinyin affords a convenient tool for improving the future of language learning research by building on the parameters of traditional AGL or existing miniature language paradigms.

Keywords Modified miniature language · Artificial grammar learning · Language learning · Sentence processing · Mini Pinyin

Introduction

Language learning is a dynamic process that involves the extraction of meaning from linguistic elements spanning multiple scales of complexity (de Diego-Balaguer, Fuentemilla, & Rodriguez-Fornells, 2010). While the neural mechanisms subserving language learning at the syllable and word levels have been well characterised (for a review, see Davis & Gaskell, 2009), uncovering the mechanisms of higher-order language learning (e.g., sentence-level combinatorics) has proved more challenging. This has been due in part to the inherent difficulty of studying language learning in a controlled environment while preserving the ecological validity of the paradigms utilised. These challenges are further exacerbated by cross-linguistic differences in the use of cues for sentence interpretation, including word order, case-marking and animacy (MacWhinney, Bates, & Kliegl, 1984).

In this paper, we present and validate a new modified miniature language paradigm modelled on Mandarin Chinese. This paradigm was designed to study higher-order language learning from a cross-linguistic perspective. First, we briefly summarise previous research on higher-order language learning, including traditional artificial grammar learning paradigms. We then introduce the concept of modified miniature languages, and the insights they have afforded into the mechanisms subserving sentence-level processing. After highlighting the strengths and limitations of current paradigms, we

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-020-01473-6) contains supplementary material, which is available to authorized users.

✉ Zachariah R. Cross
Zachariah.Cross@unisa.edu.au

[1] Cognitive and Systems Neuroscience Research Hub, University of South Australia, Adelaide, Australia

[2] Body in Mind Research Group, The Sansom Institute for Health Research, University of South Australia, Adelaide, Australia

introduce our miniature language and present behavioural data to support its use in future psycholinguistic and neurolinguistic research.

Artificial grammar learning: Simulating linguistic complexity

Early theories of implicit learning (Hayes & Broadbent, 1988; Reber, 1976; Reber, Kassin, Lewis, & Cantor, 1980) proposed that the encoding and generalisation of complex structural knowledge is learned more efficiently implicitly. Indeed, evidence provided by Reber (1976) demonstrated that individuals are capable of acquiring grammatical knowledge by memorising exemplars without any explicit instructions regarding the underlying rules. This observation in part inspired the development of artificial grammar learning (AGL) paradigms, which aim to emulate the structural aspects of language acquisition in real time (Kepinska, Pereda, Caspers, & Schiller, 2017; Wilson et al., 2013). AGL paradigms consist of strings of stimuli instantiating a grammar that is implicitly learned, followed by a test phase where novel strings are classified as grammatical or ungrammatical (Petersson, Folia, & Hagoort, 2012). Unlike natural languages, most AGL paradigms do not contain semantic, phonological and pragmatic properties, thus enabling the experimenter to control for these various linguistic elements, while also controlling for prior (language) learning (Folia et al., 2010).

AGL paradigms have been used to investigate domain-general mechanisms of language acquisition (e.g., statistical learning ability), and to compare cross-species language acquisition. This work has revealed similarities and differences in human and non-human primates' abilities to acquire complex, hierarchical sequences (Milne et al., 2016; Mueller, Milne, & Mannel, 2018; Wilson et al., 2013; see Fig. 1 for a summary of behavioural findings from select AGL studies). For example, rhesus monkeys (*Macaca mulatta*) show analogous event-related potential (ERP) modulations for the violation of non-adjacent syllable sequences to those observed in early human developmental stages (Milne et al., 2016). Further, adult humans are capable of acquiring multiple non-adjacent dependencies (e.g., sequence indices of $[A_1[A_2[A_{3,}B_1]B_2]B_3]$) that are argued to occur in natural language, such as in gender agreements between non-adjacent nouns and verbs (for a review of the relation between AGL and language processing, see Uddén & Männel, 2018). Adjacent and non-adjacent sequences have also been shown to be learned to the same degree (e.g., Uddén, Ingvar, Hagoort, & Petersson, 2012), with Broca's area playing an important role in the processing of complex, structured sequences (Uddén, Ingvar, Hagoort, & Petersson, 2017).

One prominent artificial grammar (e.g., Friederici, Steinhauer, & Pfeifer, 2002; Kepinska et al., 2017) that has been used to investigate the mechanisms underlying grammar learning is BROCANTO. BROCANTO consists of different word classes (e.g., nouns, verbs, noun and verb modifiers) and follows a subject-verb-object (SVO) sentence structure (for a full description, see Friederici et al., 2002). Recent research has demonstrated that individuals with high language analytical ability (LAA) recruit different neural networks during the learning and processing of BROCANTO compared to their low-LAA counterparts[1] (Kepinska, de Rover, Caspers, & Schiller, 2017b). Behaviourally, those with high LAA are more sensitive to grammatical violations, while stronger functional connectivity within working memory, visual, cerebellar and emotion-related networks predicted successful language learning. Subsequent work revealed that high task performance predicts increased functional connectivity of Broca's area and bilateral hippocampi (Kepinska, de Rover, Caspers, & Schiller, 2017a). In a similar EEG study (Kepinska, Pereda, et al., 2017), low-frequency (i.e., < 12 Hz) phase synchrony decreased linearly, while high-frequency (i.e., > 15 Hz) activity increased across the learning phase. Together, these findings demonstrate that grammar learning is a complex, multifactorial phenomenon that draws upon dynamic anatomical and neurophysiological mechanisms, and that this process is modulated by individual differences in LAA. They also highlight that BROCANTO has provided a useful window onto the neurobiology of grammar learning; however, given that BROCANTO is an artificially constructed language with – typically – no associated meaning, it is not possible to compare behavioural and neural findings with those reported in native speakers. Further, differences in language learning may be explained by more domain-general mechanisms, such as statistical learning ability (Erickson & Thiessen, 2015; Frost, Armstrong, & Christiansen, 2019; Jost & Christiansen, 2017), over and above that of LAA.

While this body of research has thus provided important insights into the (non-) human capacity to learn complex sequences (see Table S1 in the supplementary material for a detailed summary of findings from select higher-order language learning studies), many AGL paradigms have not fully captured the complexity of higher-order language learning, given that many studies do not provide semantic or contextual information to learners (e.g., Mueller et al., 2014). However, more recent AGL studies have shown that semantic biases facilitate the learning of simple artificial grammatical rules (e.g., Poletiek & Lai, 2012), while cues marking boundaries of major units in a sequence are learned more efficiently when indicated by prosodic information (Mueller, Bahlmann, &

[1] Language learning aptitude is typically defined as the fixed ability of an individual to acquire a (second) language and is often operationalised through standardised tests that measure rote learning, grammatical sensitivity and inductive language learning ability (Kepinska et al., 2017). Language analytic ability, a component of language learning aptitude, is defined as "the ability to infer linguistic systematicities from the input and make generalizations" (Roehr-Brackin & Tellier, 2019, p. 2).

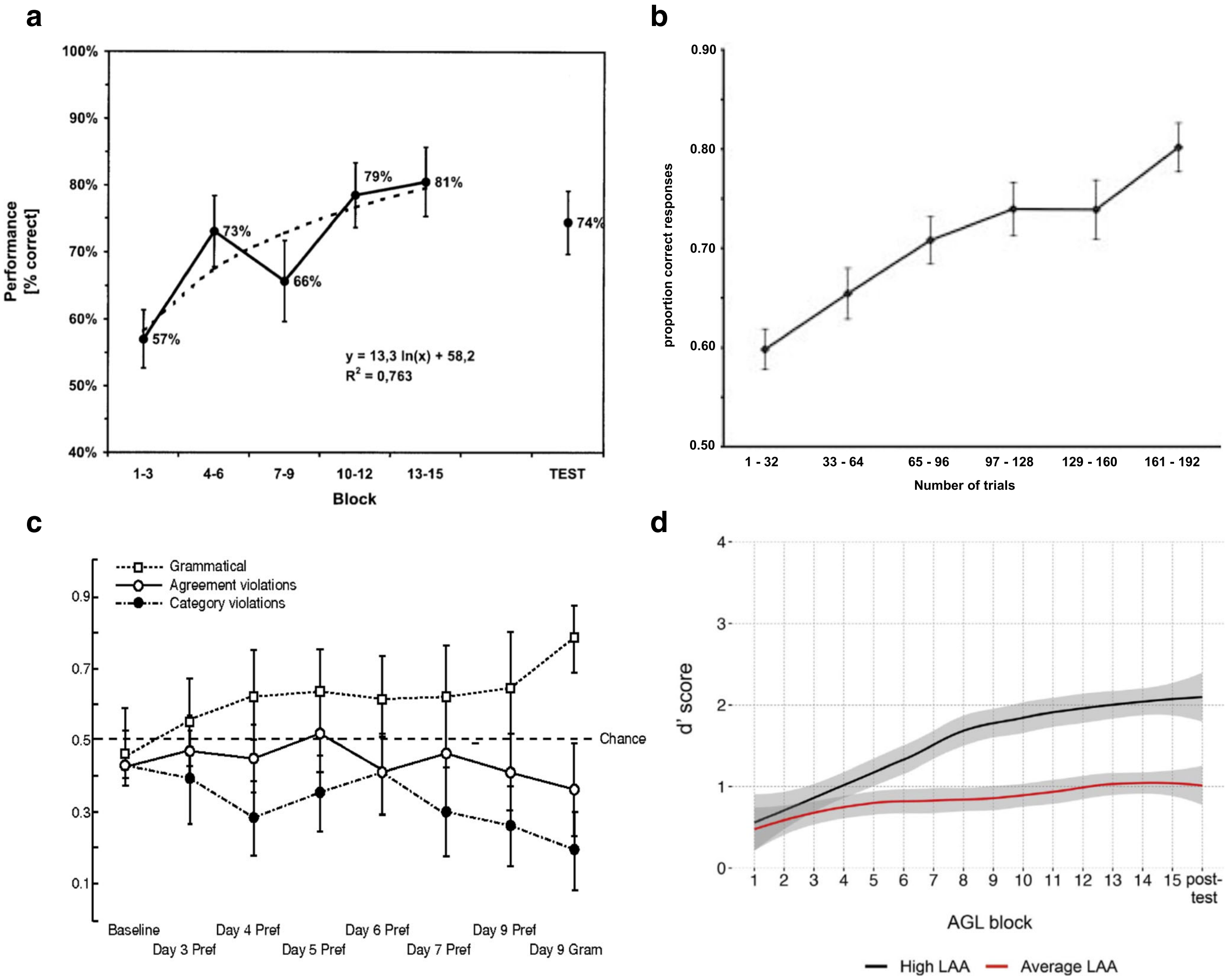

Fig. 1 Behavioural performance on grammaticality judgement tasks across time and exposure to various AGL paradigms. These judgement tasks require participants to identify illegal from legal strings of stimuli based on the rules acquired during a learning phase. (**a**) Percent correct on judgement task across experimental block reported in Opitz and Friederici (2003); (**b**) proportion of correct responses across task trials reported in Opitz and Friederici (2007); (**c**) proportion of correct responses over a nine-day exposure period reported in Folia et al. (2010); (**d**) d' scores on a grammaticality judgement task across experimental blocks between participants with a high and average language analytical ability (Kepinska et al., 2017). Permission to reuse images was obtained from the copyright holder via RightsLink®

Friederici, 2010). Further, a study using BROCANTO (Friederici, Steinhauer, & Pfeifer, 2002) had participants learn the meaning of noun and verb phrases by playing a board game that contained verbal descriptions of the stimuli. Here, grammatical violations elicited an early negativity and late positive ERP component for sentential rules not present in participants' native language. Taken together, these studies indicate that successful higher-order language learning relies on both grammatical and non-grammatical cues (e.g., semantics, prosody); however, many AGL studies train participants to reach a high level of proficiency on grammatical knowledge prior to the testing phase (e.g., 95% accuracy criterion; Friederici et al., 2002; Mueller, Hirotani, & Friederici, 2007), making it difficult to characterise the mechanisms underlying earlier stages of language learning.

In related attempts to extend beyond the learning of structural regularities, studies have used stimuli based on existing languages (Mueller, Hirotani, & Friederici, 2007; Mueller et al., 2014; Mueller, Hahne, Fujii, & Friederici, 2005), otherwise known as modified miniature languages (MML).

Modified miniature languages as ecologically valid models for studying higher-order language learning

MMLs contain a set of words belonging to syntactic categories that are combined into sentences based on grammatical regularities adapted from the language on which the MML is based (Mueller, 2006). In MMLs, participants often learn the meaning of individual words via picture-word pairs, and then complete a recognition memory test of word meaning.

Participants with accuracy scores above a set threshold are then exposed to a sentence learning phase, in which they are presented with grammatical picture-sentence combinations (Mueller et al., 2007). After a delay period, participants are tested on their ability to discriminate grammatical from ungrammatical sentences.

In contrast to BROCANTO's artificial design, Mini-Nihongo (Mueller, 2006; Mueller et al., 2007) is an MML modelled on Japanese. This MML contains four nouns, four verbs, three postpositions, two numeral classifiers, two numerals, one adjective and one temporal adverb, and contains SOV and OSV word orders (for a full description of Mini-Nihongo, see Mueller, 2006). An example sentence from Mini-Nihongo is given in (1).

(1) Ni hiki no neko ga ichi wa no hato o tsukitobasu tokoro desu.

[small-animal] [gen.] cat [nom.] 1 [bird] [gen.] pigeon [acc.] push
away take place.
"*Two cats are pushing away one pigeon*."

Mini-Nihongo has primarily been used to assess auditory language learning in native Japanese and German speakers (Mueller, 2006; Mueller et al., 2007). Across these studies, rule violations elicited an N400-P600 biphasic ERP response in native Japanese speakers, while non-native learners demonstrated a frontally distributed negativity and "native-like" late-positivity. Importantly, Mini-Nihongo contains many of the linguistic features relevant for sentence interpretation, including case-marking and animacy; however, while it includes both subject-before-object and object-before-subject orders, basic constituent order was always fixed as verb-final. In addition, Japanese and German share a relatively free word order, which heavily influences role assignment (e.g., Bornkessel-Schlesewsky et al., 2011; MacWhinney et al., 1984). This is an important point, since the assignment of thematic roles to noun phrases varies between languages (Bates, Devescovi, & Wulfeck, 2001; Bornkessel-Schlesewsky et al., 2011; MacWhinney et al., 1984). Native English speakers typically interpret the first noun as the Actor (the active, controlling participant) and the second noun as the Undergoer (the affected participant), irrespective of semantic cues (MacWhinney et al., 1984). By contrast, in languages like German and Japanese, thematic role assignment is based strongly on other cues, such as case marking and animacy (see Table 1 for sentence examples from languages that rely on sequence-dependent and sequence-independent processing strategies).

As illustrated in Table 1, in German and Turkish, dependencies (role assignment) can be indicated by accusative and nominative case marking (Bornkessel-Schlesewsky et al., 2015). Conversely, in languages like English and Dutch, animacy and case marking are overridden by word order.[2] For instance, *the javelin has thrown the athletes* can only be interpreted as the javelin (Actor) threw the athletes (Undergoer), demonstrating that argument position governs interpretation (Bornkessel & Schlesewsky, 2006; Bornkessel-Schlesewsky et al., 2011). Such differences in role assignment strategies based on different word orders have not been independently assessed in existing MML paradigms, limiting the generalisability of results to typologically diverse languages. Further, Japanese shares a number of linguistic properties with German (i.e., flexible word order, morphological case-marking). Thus, it is unknown whether the consolidation of newly acquired linguistic knowledge of Mini-Nihongo was supported by native German speakers' pre-existing language-related schemas (Cross, Kohler, Schlesewsky, Gaskell, & Bornkessel-Schlesewsky, 2018; Mirkovic & Gaskell, 2016). In the general memory literature, it is well known that prior knowledge and systematicity influence the consolidation and generalisation of newly acquired associations (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Gilboa & Marlatte, 2017; Mirkovic & Gaskell, 2016). If new information is consistent with existing knowledge, then less time is required for consolidation. From this perspective, effects of systematicity and prior knowledge could be tested by having participants learn a grammar that includes rules that deviate from those of their native language. If performance is higher for rules that are consistent with learners' native language, then this would constitute evidence that new linguistic information is acquired more easily by networks subserving native language encoding.

Finally, while AGL paradigms have assessed the influence of more domain-general mechanisms (e.g., statistical learning ability; Rohrmeier & Cross, 2014) on the learning of structural regularities, MML studies have studied higher-order language in the absence of these factors. Of the mechanisms posited to underlie sentence-level learning, statistical learning has proved to be particularly powerful in predicting individual differences in language learning ability (Misyak & Christiansen, 2012; Misyak, Christiansen, & Tomblin, 2010; Romberg & Saffran, 2010). Statistical learning builds on the assumption that the brain is fundamentally engaged in the task of extracting the statistical regularities of its environment in order to generate accurate predictions about the way external states are likely to unfold across multiple spatiotemporal scales (Friston, 2010, p. 201). This perspective is especially suited to language, a domain where information is represented across multiple hierarchical levels (e.g., phonemes, words, sentences). For instance, during language learning, the transitional probability between syllables is extracted in order to encode associations between word form and meaning (for

[2] In addition, note that both of these languages only show remnants of a case-marking system (e.g., in personal pronouns such as *I* vs. *me*).

Table 1. Sentence examples from languages that rely on sequence-dependent and sequence-independent processing strategies

Language	Sentence example(s)	Processing strategy
English	The girl hit the boy	Sequence Dependent
Turkish	Kitabı adam okuyor *Book[ACC] man read.* *"The man is reading the book."*	Sequence Independent
German	Den Jungen küsste das Mädchen *The boy[ACC] kissed the girl[NOM]* *"The girl kissed the boy."*	Sequence Independent
Dutch	De speer heeft de atleten geworpen *The javelin has the athletes thrown* *"The javelin has thrown the athletes."*	Sequence Dependent

Note: Examples adapted from Bornkessel-Schlesewsky et al. (2011) and Bornkessel-Schlesewsky, Schlesewsky, Small and Rauschecker (2015). Abbreviations: NOM = nominative; ACC = accusative

review, see Friederici, 2005). By extension, during sentence processing, words are combined in order to express complex meaning, which is driven by various informational sources (i.e., word order and animacy) that are weighted according to the conditional probabilities of the language (Bates et al., 2001; Bornkessel-Schlesewsky et al., 2011; MacWhinney et al., 1984). Controlling for statistical learning ability might therefore provide a more fine-grained insight into the mechanisms subserving higher-order language learning, with higher statistical learning ability associated with a greater language learning capacity. However, given the sequence-based nature of statistical learning tasks, it is unknown whether statistical learning ability only predicts the extraction of sequence-based rules in a sentence, or also explains individual differences in relational processing between non-adjacent elements in a newly learned language.

Here, we present an MML paradigm – termed Mini Pinyin – that aims to deepen our understanding of the mechanisms subserving higher-order language learning by: (1) assessing the influence of prior (linguistic) knowledge on the consolidation of newly acquired information; (2) including linguistic elements (e.g., word order) that better reflect how languages fundamentally differ in regard to information sources relevant for sentence interpretation; (3) measuring changes in performance over the course of the experiment to better characterise the initial stages of language learning, rather than only including participants who attain a high accuracy criterion; and (4) controlling for individual differences in more domain-general mechanisms, such as statistical learning ability.

The present study

The purpose of the current study was to validate Mini Pinyin, make the material (stimuli, experimental tasks, raw data) open to the scientific community, and to build upon and complement existing paradigms used to study higher-order language learning. In particular, Mini Pinyin builds upon existing paradigms by requiring participants to (a) learn that some sentence constructions allow for a free word order, while others do not; (b) differentiate between sentences that contain strict argument versus verb position rules; and (c) learn that there are changes in cue importance from one construction to the next, which tests components of the competition model (i.e., participants need to learn to focus on different cues for interpretation; Bates & MacWhinney, 1989; Bates, McNew, McNew, MacWhinney, Devescovi, & Smith, 1982). Seventy-six monolingual native English speakers learned Mini Pinyin by viewing picture-sentence pairs before completing a grammaticality judgement task and a separate visual statistical learning task. Generalised linear mixed-effects modelling of grammaticality judgements was used to examine language learning. We hypothesised that (1) the probability of a correct response would increase across the duration of the judgement task; (2) consistent with principles of prior knowledge and systematicity, native English speakers would demonstrate a steeper learning curve for sentences that follow the canonical English SVO word order relative to sentences with verb-final constructions; and (3) statistical learning ability would be positively associated with performance on the sentence judgement task.

Method

Participants

Participants included 76 healthy, monolingual, native English-speaking adults (27 female) ranging from 18 to 40 years old ($M = 24.9$, $SD = 6.78$). Participants reported having never been exposed to Mandarin Chinese. All participants reported normal or corrected-to-normal vision and had no current or past psychiatric conditions or intellectual impairment.

One participant was excluded from analysis on account of their button presses not being registered during the judgement task, resulting in a final sample of 75 (mean age = 24.9, *SD* = 7.07; 26 female). Note that this sample was pooled from two experiments: an electroencephalographic experiment (Cross et al., 2020; *n* = 36) and a separate behavioural experiment (*n* = 40). Both experiments were conducted at the Cognitive and Systems Neuroscience Research Hub at the University of South Australia (ethics approval number: 201496).

Control measures

A visual statistical learning task (for a detailed description, see Siegelman, Bogaerts and Frost, 2017a) was administered to participants, as statistical learning ability has been shown to predict individual differences in language learning (Daltrozzo et al., 2017). Briefly, this task contained 16 visual shapes and included a familiarisation phase followed by a test phase. Prior to the familiarisation phase, the 16 shapes were randomly organised for each participant into a set of eight triplets. During familiarisation, the eight triplets appeared one at a time in a random order for a total of 24 trials. Each shape appeared for 800 ms, followed by a 200 ms inter-stimulus interval (ISI). The test phase was divided into two blocks: (1) 34 pattern recognition items and (2) eight pattern completion items. Participants were required to choose the correct answer among a set of foil items using a keyboard, with the total score on the task ranging from 0 to 42 based on the number of correct responses. Figure 2 illustrates the blocks of the recognition phase. As an additional control, the Stanford Sleepiness Scale (SSS; Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973**)** was completed by participants at the beginning and end of the experiment in order to control for self-perceived sleepiness.

Mandarin Chinese as a suitable language model

Mandarin was chosen as it allows for a comparison of sequence-based and dependency-based combinatorics using word order restrictions and classifiers (Bornkessel-Schlesewsky et al., 2011). In languages that rely on order-based processing strategies such as English, the first noun phrase (NP) is typically interpreted as the Actor. For example, the sentence *the apple ate the boy* can only yield an implausible interpretation, as the apple is assigned the role of Actor and the boy as Undergoer. Conversely, in Mandarin, word order is flexible and is based more on contextual cues. As such, a plausible interpretation of the previous example would be *the boy ate the apple*, as role assignment is heavily influenced by semantic cues (e.g., animacy; Li, Bates, & MacWhinney, 1993; Wang, Schlesewsky, Bickel, & Bornkessel-Schlesewsky, 2009). However, in Mandarin, word order can also become fixed: when a coverb (i.e., *bǎ*/*bèi*) precedes the second NP, comprehenders are required to interpret the sentence based on the word order, rather than the animacy status of the NPs (Bornkessel-Schlesewsky et al., 2011; see Table 2 for sentence examples in Mandarin that include *bǎ* and *bèi*)[3].

As shown in Table 2, *bǎ* and *bèi* modulate the positioning of the NPs: *bǎ* indicates the direction of the action from the first NP to the second NP (that is, it renders the first NP as Actor and the second NP as the Undergoer), *bèi* reverses this pattern. In addition to including both fixed and flexible word orders, Mandarin also contains classifiers, the function of which is to group nouns into specific categories, enabling quantification (for more information on Mandarin classifiers, please see Gao & Malt, 2009; Her, Chen, & Yen, 2017; Zhang, 2007) . Mandarin classifiers belong to five main categories: (1) group, (2) container, (3) standard measure, (4) temporary, and (5) individual classifiers. By and large, any noun that denotes a countable object requires a classifier. Further, as classifiers are category-specific (e.g., the classifier *zhi* can only be used to quantify stick-like objects, while *ben* is used to quantify books), they are likely shaped by mechanisms of associative memory that are critical for accurate sentence comprehension. From this perspective, classifiers provide a useful basis for characterising the associative memory mechanisms underlying sentence-level processing.

Vocabulary and structure of Mini Pinyin

Mini Pinyin contains 16 transitive verbs, 25 nouns, 2 coverbs and 4 classifiers. The nouns were subdivided into 10 human entities, 10 animals and 5 objects (see Table 3 for a summary of the vocabulary of Mini Pinyin). Each category of noun was associated with a specific classifier, which preceded each NP in a sentence. As described in Table 3, *ge* specifies a human noun, *zhi* for animals, and *xi* and *da* for small and large objects, respectively.

In regard to the structural constraints of the sentences, Mini Pinyin includes two types of manipulations. The first manipulation involves altering the consistency rule of the “classifier-noun” pairs. For example, in grammatical sentences, classifiers are consistent with their associated NP (e.g., ge [human]-chushi [chef]). By contrast, in ungrammatical sentences, classifier-noun pairs are violated (e.g., zhi [animal]-piqiu [ball]). The second manipulation is based on word order rules: there are four possible word order variations, involving both sequence- and dependency-based interpretation strategies. Sentences that are dependency-based are illustrated in (1), while sequence-based sentences are illustrated in (2):

[3] Coverbs are verbs which came to be used as prepositions through a process of grammaticalisation (Hopper & Traugott, 2003; for a discussion in a psycholinguistic context, see Bornkessel-Schlesewsky et al., 2011).

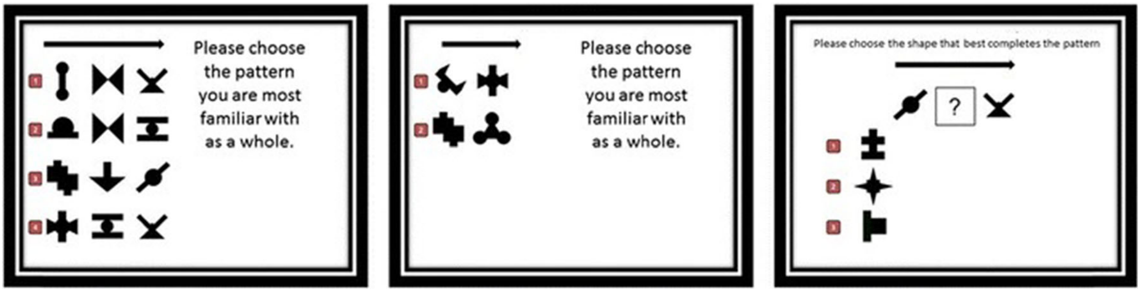
Please choose the pattern you are most familiar with as a whole.
Please choose the pattern you are most familiar with as a whole.
Please choose the shape that best completes the pattern
?

Fig. 2 Example of three trials from the test (left to right): (1) four-forced-choice pattern recognition trial with triplets; (2) two-forced-choice pattern recognition trial with pairs; and (3) pattern completion trial for a triplet. Reproduced with permission from Siegelman et al. (2017a).

(1)
(a) xi shubao xile ge faguan.
(small object) bag wash (human) judge.
“he judge washes the small bag.”
(b) ge hushi zhaole da shubao.
(human) nurse photograph (large object) bag.
“The nurse photographs the large bag.”
(2)
(a) zhi junma *bặ* xi pingguo chile.
(animal) horse *bặ* (small object) apple eat.
“he horse eats the small apple.”
(b) da shubao *bèi* ge yisheng dale.
(large object) bag *bèi* (human) doctor hit.
“The doctor hits the large bag.”

As is apparent from (1a), sentences that do not contain a coverb (*bặ*/*bèi*) yield a flexible word order, such that understanding *who is doing what to whom* is not dependent on the ordering of the NPs. Instead, determining *who is doing what to whom* is facilitated by animacy cues, such as in (1a), where despite the first NP being the bag, the judge is interpreted as the Actor, given that it is implausible for a bag to actively wash a human. By contrast, sentences such as (2a) yield a fixed word order, such that the inclusion of *bặ* or *bèi* renders the first NP either the Actor or Undergoer, respectively. Note that the positioning of the verb is critical in sentences with and without coverbs. With coverbs, the verb must be placed at the end of the sentence, while in constructions without coverbs, the verb must be positioned between the NPs.

Table 2 Examples of *bǎ* and *bèi* sentence structures (examples from Wang, Schlesewsky, Philipp, & Bornkessel-Schlesewsky, 2012)

Coverb	Sentence examples
BÈI-PLAUSIBLE	Zhēntàn bèi zǐdàn jí zhòng *Detective BÈI bullet hit.* *“The detective was hit by the bullet.”*
BÈI-IMPLAUSIBLE	Zhēntàn bèi bǎochí zǐdàn *Detective BÈI bullet kept.* *“The detective was kept by the bullet.”*
BǍ-PLAUSIBLE	Zhēntàn bặ tíng de zǐdàn *Detective BǍ bullet kept.* “The detective kept the bullet.”
BǍ- IMPLAUSIBLE	Zhēntàn bặ zǐdàn jí zhòng *Detective BǍ bullet hit.* *“The detective hit the bullet.”*

Based on the above manipulations, the use of Mini Pinyin permits the ability to measure responses to incorrect classifier noun pairs and a number of incorrect word orders. These manipulations aim to measure aspects of language-related associative memory and sequence processing via the classifier-noun pairs and word order rules, respectively.

Experimental protocol and procedure

Participants learned the 25 nouns prior to the main experimental session to ensure that they had a basic vocabulary of nouns in order to successfully learn the 32 transitive verbs (see Fig. 4 for a schematic of the vocabulary booklet). Participants collected a paired picture-word vocabulary booklet containing the 25 nouns and were instructed to learn the meaning of the words. They were required to use an activity log in which they recorded when they studied the vocabulary booklet. After a minimum of three days of vocabulary learning, participants returned to complete the main experimental session. Note that participants only learned the nouns explicitly, while other elements in the language (i.e., classifiers, verbs, coverbs and grammar) were learned during the sentence learning phase, as described below.

Vocabulary test

Prior to the main experimental session, a lab-based vocabulary test was administered in order to ensure participants had successfully learned the nouns. During the vocabulary test, participants translated the nouns from Mini Pinyin into English using a keyboard. Each trial began with a 600-ms fixation cross, followed by the visual presentation of the noun word form. Participants had 20 s to respond to each noun, and only participants who scored >84% – corresponding to 21/25 correct responses – were eligible to complete the main experiment. The proportion of individuals who did not pass the

Table 3 Full vocabulary of Mini Pinyin

Verbs		Nouns			Classifiers	Coverb
No object	Object	Animal	Human	Object		
Tianle (to lick)	Xile (to wash)	Maomi (cat)	Chushi (chef)	Shubao (bag)	Da (big)	Bă
Zhuile (to chase)	Leangle (to measure)	Junma (horse)	Haidao (pirate)	Xianjiao (banana)	Xi (small)	Bèi
Nale (to hold)	Kanle (to observe)	Yegou (dog)	Jingcha (policeman)	Shuben (book)	Ge (human)	
Chile (to eat)	Zhaole (to photograph)	Tuzi (rabbit)	Hushi (nurse)	Pingguo (apple)	Zhi (animal)	
Rengle (to throw)	Shele (to shoot)	Laohu (tiger)	Yisheng (doctor)	Piqiu (ball)		
Tile (to kick)	Dale (to hit)	Houzi (monkey)	Faguan (judge)			
Beile (to carry)	Kunle (to tie)	Xiong (bear)	Feixing (pilot)			
Shuale (to swipe)	Zhuole (to capture)	Laoshu (rat)	Xiaofang (firefighter)			
		Daxiang (elephant)	Shuishou (sailor)			
		Shizi (lion)	Niuzai (cowboy)			

vocabulary test was small (e.g., approximately less than 10 cases); however, the exact number was not recorded. As such, all 76 participants included in subsequent analyses obtained over 84% correct on the vocabulary test. See Fig. 3 for a schematic of the vocabulary booklet.

Sentence learning

Grammatical picture-sentence pairs were presented to participants using OpenSesame Software (Mathot, Schreij, & Theeuwes, 2012). During sentence learning, pictures were used to express events occurring between two entities. While participants were aware that they were to complete sentence judgement tasks following the learning phase, no explicit feedback was given during the learning task.

Sentences were constructed using the learned NPs and novel verbs, coverbs and classifiers. As Mini Pinyin contains a flexible word order, each picture corresponded to four sentence variations with either Actor-first or Undergoer-first orders with varying verb positions, fixed via the addition of the coverbs *bặ* or *bèi*, respectively. Participants were presented with a fixation cross for 2000 ms, followed by the picture illustrating the event between two entities for 4000 ms. A sentence describing the event in the picture was then presented on a word-by-word basis. Each word was presented for 700 ms followed by a 200 ms inter-stimulus interval (ISI). This pattern continued for the 128 sentence-picture combinations until the end of the task, which took approximately 40 minutes (including three self-paced breaks). Each of the four grammatical sentence constructions illustrated in Fig. 2 were presented equally (i.e., 32 trials each); however, stimuli were pseudo-randomised, such that no stimuli of the same construction were presented consecutively. On completion of the sentence learning session, participants completed the sentence judgement task (see Fig. 5a for a schematic of the learning task).

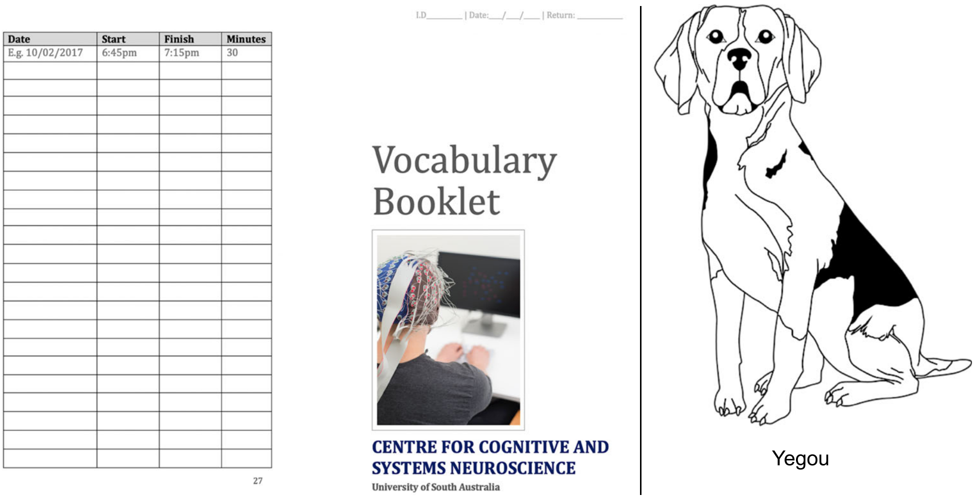
Date
Start
Finish
Minutes
E.g. 10/02/2017
6:45pm
7:15pm
30
27
I.D_______ | Date:__/__/__ | Return: _______
Vocabulary Booklet
CENTRE FOR COGNITIVE AND SYSTEMS NEUROSCIENCE
University of South Australia
Yegou

Fig. 3 Illustration of the vocabulary booklet. Participants were required to learn the meaning of the 25 nouns (e.g., *yegou* = dog) prior to the experimental session

Judgement task

During the sentence judgement task, 288 novel sentences without pictures were presented word-by-word with a presentation time of 600 ms and an ISI of 200 ms. Participants received feedback after their response to facilitate learning. The sentences in the judgement task also included 32 novel verbs with the same ending as learned verbs (i.e., *le*). Hence, the interpretation of these sentences is based on the grammatical rules of the language, rather than the meaning of the main verb phrase. Fig. 4b illustrates the sequence of events in the judgement task.

Participants were instructed to read all sentences attentively and to judge their grammaticality (yes/no) via a button press on a standard keyboard. As a cue for judgment, a question mark appeared in the centre of the computer monitor for 4000 ms after the offset of the last word, followed by feedback that indicated whether participants' responses were correct or incorrect. Two lists of sentence stimuli were created, which were counterbalanced across participants. One hundred forty-four of the sentences were grammatical, with each of the four grammatical constructions shown an equal number of times. The remaining 144 sentences were ungrammatical constructions, violating either the position of the verb, the position of the Actor/Undergoer in *bǎ*/*bèi* constructions, and noun-classifier pairings. Stimuli were pseudo-randomised, such that no same stimuli followed each other.

Main experimental protocol

Participants completed the vocabulary test before completing the sentence learning and sentence judgement tasks. Participants who scored below 84% accuracy on the vocabulary test were not eligible to continue with the experiment and received a $10 honorarium for their time. Participants completed the statistical learning task at the end of the experimental session. See Fig. 5 for an illustration of the experimental protocol.

Data analysis

Three measures were calculated from performance on the judgement task: (1) grammaticality judgments calculated on a trial-by-trial basis, determined by whether participants correctly identified grammatical and ungrammatical sentences; (2) the response time of grammaticality ratings in milliseconds derived from the judgement task; and (3) based on signal detection theory (Stanislaw & Todorov, 1999), hit rate (HR) and false alarm rate (FA) were computed to derive the discrimination index (d'), defined as the difference between the z transformed probabilities of HR and FA (i.e., d' = z[HR] − z[FA]).

Data were analysed in *R* (R Core Team, 2020) using generalised linear mixed-effects logit models (GLMM; Bates, Maechler, Bolker, & Walker, 2015; Jaeger, 2008) fit by maximum likelihood using the *lme4* (Bates, 2010; Barr, 2013) and *glmm* (Bates et al., 2015) packages. Generalised (logit) mixed-effects models are an appropriate method for analysing data from repeated measure designs, particularly in psycholinguistic research, as they account for within- and between-subject variance, as well as variance introduced by items (Baayen, Davidson, & Bates, 2008; Judd, Westfall & Kenny, 2012; Van Dongen, Olofsen, Dinges, & Maislin, 2004). Further, logit mixed models appropriately account for binomial response variables (e.g., 1 = correct, 0 = incorrect; Jaeger, 2008). As such, a logit mixed model is particularly suited to the current data due to: (a) inter-individual differences in language learning and performance on the judgement task; (2) the use of categorical response variables (i.e., binomial correct/incorrect responses) as a measure of language learning and (3) taking into account item variability, given that items may vary in familiarity across participants and thus influence learning outcomes (Baayen et al., 2008; Quené, Huub, & Bergh, 2008). Further, the use of a trial-based outcome variable in our main statistical models allows for more fine-grained analyses of by-item and by-participant variability, which are lost in aggregated variables, such as proportion correct or d'.

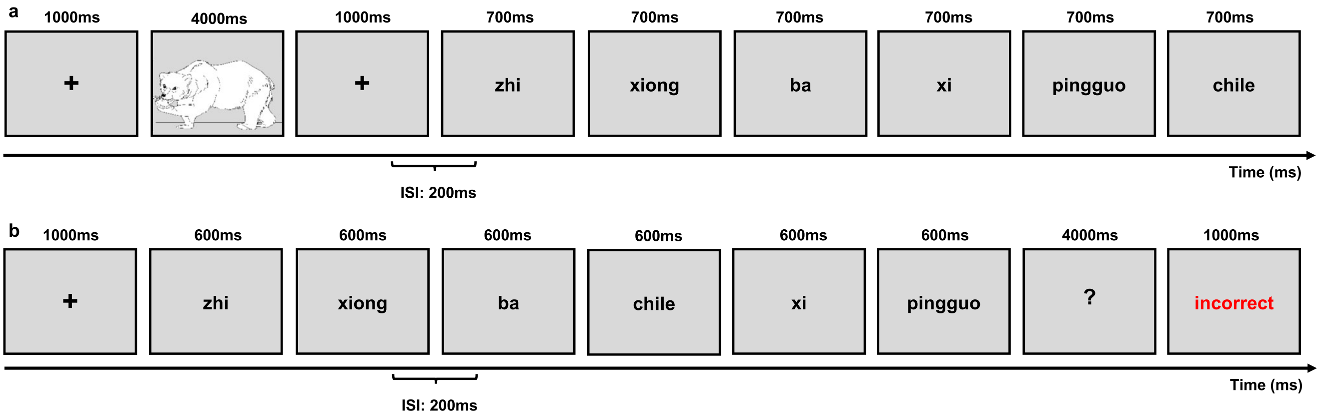

Fig. 4 (**a**) Schematic representation of the sentence learning task. (**b**) Schematic representation of the sentence judgement task

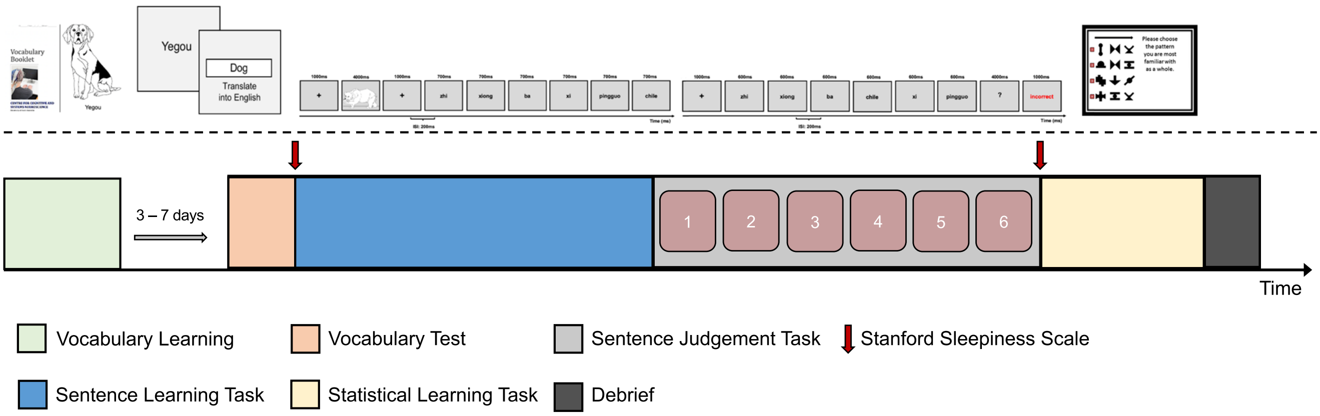

Fig. 5 Diagram representing the time course of the experimental tasks. The red numbered blocks embedded in the sentence judgement task indicate the testing blocks (1–6)

The models included fixed effects for "Block" (blocks 1–6), "Sentence Type" (fixed, flexible) and "Grammaticality" (grammatical, ungrammatical) and specified interactions for all of these factors. The random effects structure included a random intercept for participants and items. More complex random effect structures involving random slopes by participant did not converge. In order to control for potential fatigue effects, self-perceived sleepiness (measured by the SSS) was added into the GLMM as a fixed effect, given that sleepiness modulates performance on a range of cognitive tasks (Franzen, Siegle, & Buysse, 2008) including grammatical reasoning tasks (Dorrian, Lamond, & Dawson, 2000). Trial-based response accuracy was specified as the dependent variable (DV), with 1 coded as a correct response and 0 coded as an incorrect response for the 288 trials for each participant, totalling 21,600 observations. Type II Wald $\chi2$ tests from the *car* package (Fox, 2011) were used to provide p value estimations for each effect. Post-hoc comparisons for main effects were performed using the *emmeans* package (Lenth, 2020). The Holm–Bonferroni method (Holm, 1979) was used to correct for multiple comparisons. Contrasts for categorical factors in the GLMMs were coded using sum coding, which generates coefficients that reflect differences relative to the grand mean (Schad et al., 2020). Further, Block was specified as an ordered factor, and differences in trial-based response accuracy between Blocks (1–6) were examined using polynomial contrasting, which generates coefficients that reflect linear changes in the outcome variable (Narula, 1979). We adopted an 83% confidence interval (CI) threshold, which corresponds to the 5% significance level with non-overlapping estimates (Austin & Hux, 2002; MacGregor-Fors & Payton, 2013).

Beta regressions were used to assess the relationship between statistical learning ability and reaction time on d' scores. Boxplots were also used to visualise descriptive statistics of accuracy judgements for each of the sentence conditions, while a paired-sample t test was computed to determine whether self-perceived sleepiness was statistically different between the beginning and end of the experiment. Statistical significance was determined at $\alpha < .05$, while all data are presented using the mean and confidence intervals unless indicated otherwise. Effects were plotted using *ggplot2* (Wickham, 2016). Two participants did not complete the statistical learning task, while four participants did not complete the SSS; however, given that linear mixed-effects models appropriately handle missing data, these participants were not excluded from the main statistical analysis.

Results

Descriptive statistics and preliminary analyses

On average, participants attained 97.33% accuracy on the vocabulary test ($SD = 4.12$; range = 84–100%). Participants also showed a moderate degree of accuracy on the judgement task (mean accuracy range: 37–67%, also see Table 4). However, there was large inter-individual variability in performance, evidenced by the range of standard deviations for percent of correct responses (range: 15.7–24.7). The range in scores between sentence types is also visualised in Fig. 6.

On average, self-perceived sleepiness increased from the beginning ($M = 2.30$, $SD = 0.74$) to the end ($M = 3.41$, $SD = 1.39$) of the experiment, indicating that participants became increasingly fatigued across the experiment. A paired samples t test indicated that this difference was significant ($t(69) = -8.06, p < .001$, 95% CI = [−1.35, −0.81], $d = .98$), supporting the inclusion of self-perceived sleepiness as a main effect in the GLMM. Participants' performance on the statistical learning task ranged from 29% to 93% ($M = 58.38$, $SD = 16.34$).

Modelling learning across time, grammaticality and word order

To obtain an initial broad overview of participants' learning performance, we first examined how the probability of a correct response on the judgement task was modulated by Block (as a proxy for time), Grammaticality and Sentence Type (fixed or flexible word order). More fine-grained analyses targeting specific violation types will be reported in the following sections. A main effect of Block (χ2(5) = 24.90, *p* < .001) showed that the probability of a correct response increased from block 1 to block 6. Post-hoc comparisons revealed that trial-based response accuracy was significantly lower in block 1 compared to block 6 (β = −0.25, se = .06, z = −3.84, *p* = .001) and in block 3 compared to block 6 (β = −0.21, se = .06, z = −3.39, *p* = .009), indicating that performance increased across the judgement task. Further, a main effect of Type (χ2(1) = 15.08, *p* < .001) revealed that the probability of a correct response was higher for sentences with flexible versus fixed word orders. A main effect of Grammaticality (χ2(1) = 229.96, *p* < .001) demonstrated that the probability of a correct response was higher for grammatical versus ungrammatical sentences. Grammaticality also interacted with Type (Grammaticality × Type, χ2(1) = 150.24, *p* < .001), such that ungrammatical fixed word order sentences had significantly lower correct responses than grammatical fixed word order sentences, and grammatical and ungrammatical flexible word order sentences. Figure 7 illustrates this interaction effect.

We also observed a Block × Grammaticality effect (χ2(1) = 18.81, *p* = .002): the probability of a correct response for ungrammatical sentences was highest at block 4 and tapered off thereafter. In contrast, the probability of a correct response for grammatical sentences was lowest at block 3 and steadily increased until block 6. The Block × Type (χ2(5) = 4.12, *p* = .53) and Block × Type × Grammaticality (χ2(5) = 7.75, *p* = .17) interactions were nonsignificant. See S1 in the supplementary material for a full summary of the GLMM. Figure 8 resolves the significant Block × Grammaticality effect.

Table 4 Mean accuracy of grammaticality ratings by sentence type, grammaticality and word order (fixed, flexible). Standard deviations are given in parentheses

	Sentence Type	Percent correct	Reaction time (ms)
Grammatical			
Flexible	AVU	64.60 (19.49)	970.89 (330.75)
Flexible	UVA	61.28 (18.10)	953.26 (311.02)
Fixed	A*bặ*UV	65.39 (15.70)	793.01 (327.47)
Fixed	U*bèi*AV	65.65 (17.32)	769.99 (325.67)
Ungrammatical			
Fixed	A*bèi*UV	37.00 (17.84)	827.55 (351.80)
Fixed	U*bặ*AV	37.12 (17.50)	818.20 (357.31)
Fixed	A*bặ*VU	67.01 (24.75)	799.97 (356.67)
Fixed	U*bèi*VA	66.95 (24.74)	784.10 (306.76)
Flexible	AUV	64.02 (24.26)	887.52 (364.94)
Flexible	UAV	65.82 (22.06)	834.81 (323.20)
NA	NP1	50.30 (20.44)	806.90 (273.90)
NA	NP2	47.75 (18.22)	881.89 (314.14)

Note: AVU = Actor-verb-Undergoer; UVA = Undergoer-verb-Actor; AUV = Actor-Undergoer-verb; UAV = Undergoer-Actor-verb; NP1 = first noun phrase; NP2 = second noun phrase

Verb and noun phrase order rules differentially predict grammaticality judgements

Given that fixed word order sentences (i.e., *bặ* and *bèi* constructions) contain two violation types – namely Actor/Undergoer and verb position violations—we ran separate analyses examining whether the two violation types differentially influence behavioural performance. Block (1–6), Coverb (*bặ*, *bèi*), Violation (noun, verb) and Statistical Learning Ability were specified as fixed effects, with full interactions. Participant and Item were modelled as random effects on the intercept, while trial-based response accuracy was specified as the outcome variable, and self-perceived sleepiness modelled as a fixed effect to control for any fatigue-related effects. The main effect of Block was nonsignificant (χ2(5) = 10.93, *p* = .05); however, there was a significant main effect of Violation (χ2(2) = 692.10, *p* < .001). As illustrated in Fig. 9a, the probability of a correct response was significantly higher for grammatical sentences and sentences with verb position violations compared to noun position violations, suggesting that participants learned verb position rules to a higher degree than noun position rules. There was also a significant Violation × Block interaction (χ2(10) = 19.57, *p* = .03), which is resolved in Fig. 9b. Post-hoc analyses revealed that trial-based response accuracy for verb position violations was significantly lower in block 2 compared to block 4 (β = −0.50, se = .17, z = −2.93, *p* = .04). Further, when statistical learning ability was low, there was a small difference in the probability of a correct response between grammatical sentences and sentences containing noun and verb position violations. By contrast, when statistical learning ability increased, the probability of a correct response for grammatical sentences and verb position violations increased, while remaining stable for noun position violations (see S2 in the supplementary material for a full summary of the GLMM). Together, these results suggest that order-based phrase structure rules, such as verb position rules, are learned to a higher degree than dependency-based rules, such as rules governing which *type* of noun phrase (Actor or Undergoer) can occupy certain positions, and that this difference is modulated by statistical learning ability.

The influence of prior knowledge and systematicity on the consolidation of varying word orders

As discussed earlier, many MML paradigms are modelled on languages that are analogous to the languages spoken by the

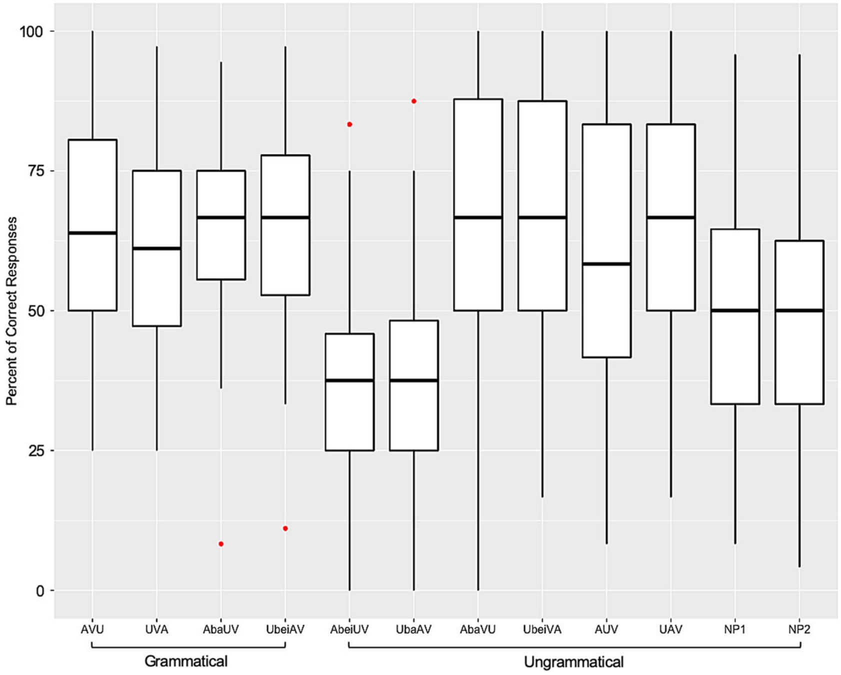

Fig. 6 Summary of percent of correct responses across all sentence conditions during the judgement task. Thick horizontal lines indicate the median; lower and upper hinges correspond to the first and third quartiles, respectively; lower and upper whiskers extend to the furthest estimate within 1.5 × interquartile range from the lower and upper hinges, respectively; red points indicate outliers

sample used. Specifically, the MML literature has hitherto not examined the possible role of similarities and differences between native language characteristics of the learners and those of the MML being learned. For example, Mini-Nihongo (Mueller, 2006) is modelled on Japanese, which shares many of the same linguistic properties as German, the language spoken by the sample in Mueller et al. (2005, 2007). Thus, it is unknown if the establishment of new memory traces of linguistic elements of the language to be learned are supported by pre-existing language-related schemas, and whether this effect is modulated by individual differences in statistical learning ability. To test this idea, we categorised the features of the sentences in Mini- Pinyin as either "similar" or "different" depending on whether their basic word order was similar to English (e.g., subject-verb-object).

Logit linked generalised linear mixed-effects models were used to determine whether our sample of native English speakers were better at encoding the structural regularities of

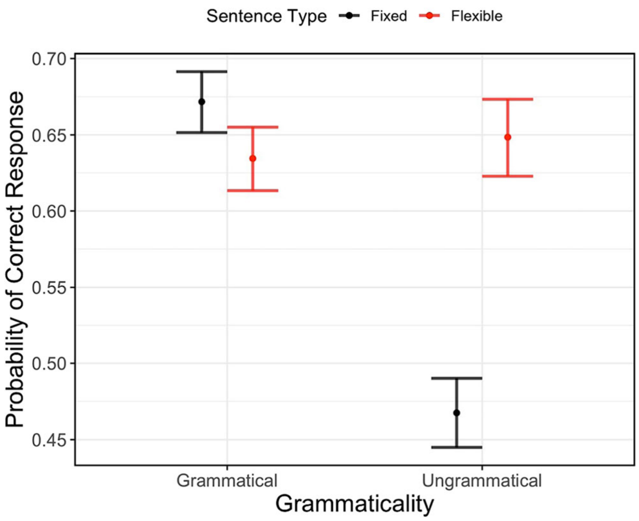

Fig. 7 Probability of correct response (*y*-axis) modelled against Grammaticality (*x*-axis; grammatical, ungrammatical) and Sentence Type. Bars represent 83% confidence intervals

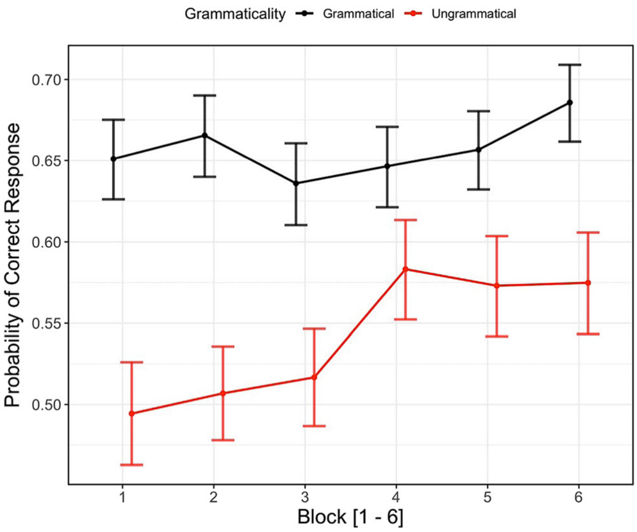

Fig. 8 Probability of correct response (*y*-axis) modelled across Block (*x*-axis; 1–6) and Grammaticality (black = grammatical, red = ungrammatical). Bars represent the 83% confidence interval

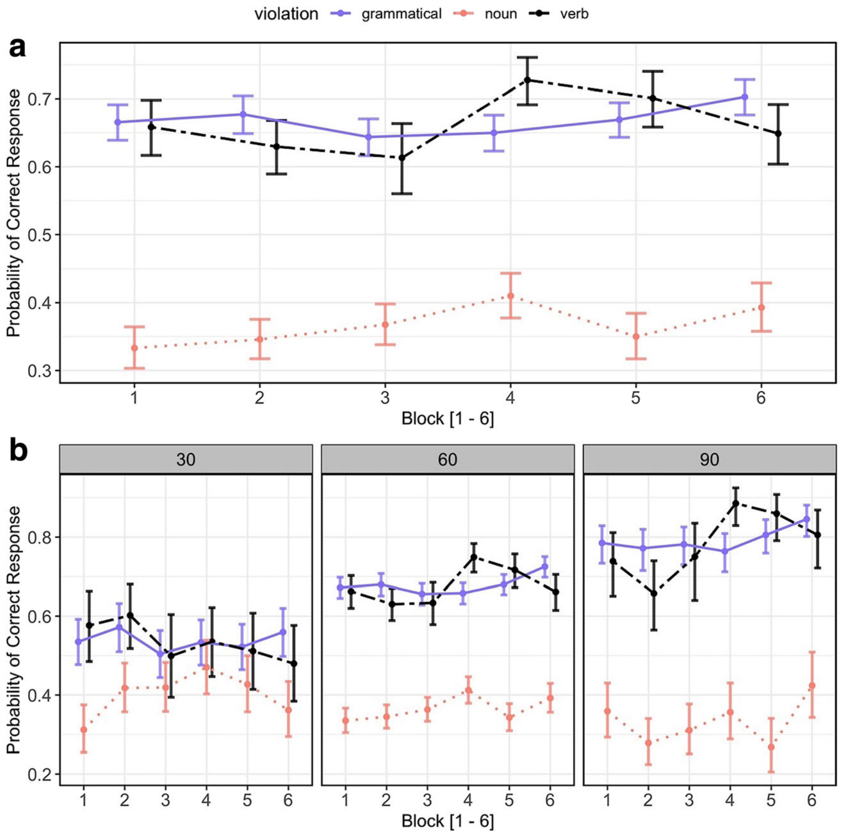

Fig. 9 Comparison in performance between noun and verb position violations for fixed word order sentences. (**a**) Probability of correct response (*y*-axis) modelled across block (*x*-axis; 1–6) and grammatical (purple solid line), noun position (orange dotted line) and verb position (black dashed line) violations. (**b**) Effect of statistical learning ability (left panel = low; right panel = high) on the probability of a correct response for grammatical and verb and noun position violations. Bars represent the 83% confidence interval

sentences with similar word orders to their native language. The model included fixed effects for "Block" (Blocks 1–6), Feature (different = verb final, similar = verb medial), Statistical Learning Ability and Grammaticality (grammatical, ungrammatical), and self-perceived sleepiness. The random effects structure included a random intercept for participants and items. Trial-based response accuracy was specified as the DV. Analyses revealed a significant effect of Block ($\chi2(5) = 21.47$, $p < .001$), with the probability of a correct response increasing across the duration of the task. Post-hoc analyses revealed that trial-based response accuracy was significantly lower in block 2 compared to block 6 for sentences similar to English ($\beta = -0.35$, se = .11, z = −3.16, $p = .02$). For sentence constructions different to English, trial-based response accuracy was significantly lower in block 3 compared to block 4 ($\beta = -0.22$, se = .06, z = −3.42, $p = .009$) and in block 3 compared to block 6 ($\beta = -0.20$, se = .06, z = −3.01, $p = .03$).

Further, the effect of Grammaticality was significant ($\chi2(1) = 247.06$, $p < .001$), indicating that the probability of a correct response was higher for grammatical sentences. The effect of Feature was also significant ($\chi2(1) = 55.59$, $p < .001$), as was the effect of Statistical Learning Ability ($\chi2(1) = 17.65$, $p < .001$). While the four-way interaction was nonsignificant, there was a significant Grammaticality × Feature × Statistical Learning Ability interaction ($\chi2(1) = 14.91$, $p < .001$). See S3 in the supplementary material for a full summary of the GLMM. The significant interaction of Grammaticality, Feature and Statistical Learning Ability is resolved in Fig. 10.

As shown in Fig. 10, the probability of a correct response was equivalent for similar and different grammatical constructions, increasing with higher statistical learning ability in each case. By contrast, performance for ungrammatical constructions differed by whether the sentence structure was similar or different to English. For constructions following a different verb position to English, participants performed at chance level independently of statistical learning ability, while ungrammatical constructions with a similar verb position to English showed a positive association with statistical learning ability in the majority of blocks.

Classifier-noun pairing violations

To assess participants' ability to detect classifier-noun pairing violations, we ran a separate logit linked generalised mixed-effects model with Block (1–6) and Violation (noun phrase one, noun phrase two) as fixed effects and trial-based response accuracy as the outcome variable. Subject and item were specified as random effects on the intercept. Analyses revealed non-significant effects of Block ($\chi2(5) = 6.05$, $p = .30$) and Violation ($\chi2(1) = 3.23$, $p = .07$) and a non-significant Block × Violation interaction ($\chi2(5) = 1.17$, $p = .94$). As is clear from Fig. 11, and although non-significant, trial-based response accuracy decreased from block 1 to 2 before increasing thereafter, and then decreasing after block 4. Further, while the probability of a correct response was slightly higher for classifier-noun violations at the first noun phrase compared to the second noun phrase, this difference was not significant (see S4 in the supplementary material for a full summary of the GLMM).

Statistical learning ability partially explains individual differences in language learning

Here, we examine whether the proportion of correct responses on the statistical learning task and reaction time (RT; ms) from the grammatical judgement task predicts differences in correct responses on the judgement task. As a sanity check, a beta regression with a logit link function (implemented using the *betareg* package; Cribari-Neto & Zeileis, 2009) was conducted to ensure that proportion of correct responses and d' scores were related. As illustrated in Fig. 12a, the performance estimators were highly related ($\chi2(1) = 2514.8$, $\beta = .21$, pseudo $R^2 = .97$, $p < .001$), indicating that participants who attained a high proportion of correct responses were also highly sensitive to the grammatical rules.

Next, linear regressions were used to examine whether statistical learning ability and RT predicted d' scores on the judgement task (scatterplots are illustrated in Fig. 12b and c). The results of the linear regressions indicated a significant positive relationship between statistical learning ability and d' scores ($\beta = .04$, $p < .001$, $R^2 = .17$), while there was no significant effect of RT on d' scores ($\beta = -.0009$, $p = .09$, $R^2 = .02$).

Discussion

The aim of this experiment was to test the utility of Mini Pinyin for studying language learning and sentence comprehension. This miniature language contains various cross-linguistic elements, including a number of word order arrangements, which participants learned in a laboratory setting. Participants demonstrated a moderate degree of accuracy on the judgement task, with the probability of a correct response increasing across time (Block 1–6); however, performance was generally higher for grammatical sentences, indicating a bias toward grammatical constructions. We also measured individual differences in statistical learning ability, and observed a correlation between language learning and statistical learning ability, which was in line with our predictions and previous work on individual differences in statistical learning ability and artificial grammar learning. Performance on the judgement task was also modulated by different word order permutations: participants were more likely to endorse grammatical word orders that were analogous to English (verb-medial) compared to constructions that differed from English (verb-final), and participants learned verb-order rules better than rules governing the position of argument roles (Actor/Undergoer). In the following, we will discuss these findings in relation to existing MML paradigms, and the relationship between statistical learning ability and higher-order language learning. We will also provide suggestions for future applications of Mini Pinyin to characterise the neurobiological basis of higher-order language learning.

Mini Pinyin as a valid paradigm to study higher-order language learning

Several modified and artificial miniature languages have been used to characterise higher-order language learning (Friederici et al., 2002; Kepinska, Pereda, et al., 2017; Mueller, 2006; Opitz & Friederici, 2003; Weber, Christiansen, Petersson,

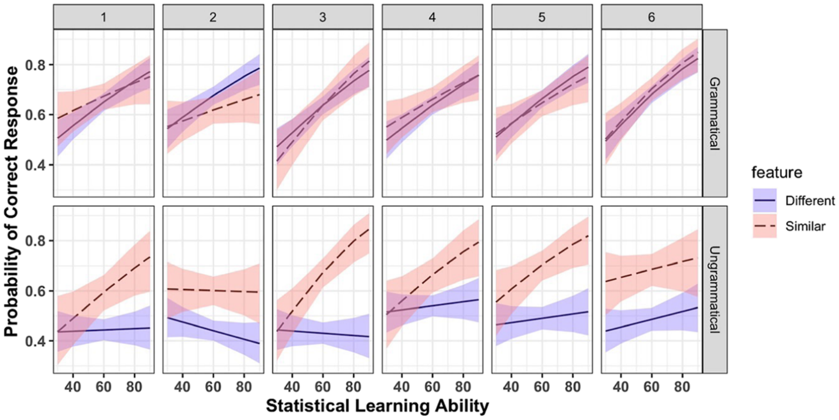

Fig. 10 Probability of correct response (*y*-axis) modelled across Block (faceted 1–6), Feature (similar = dashed line, different = solid line), Grammaticality (top = grammatical, bottom = ungrammatical) and Statistical Learning Ability (*x*-axis). Bars represent 83% confidence intervals

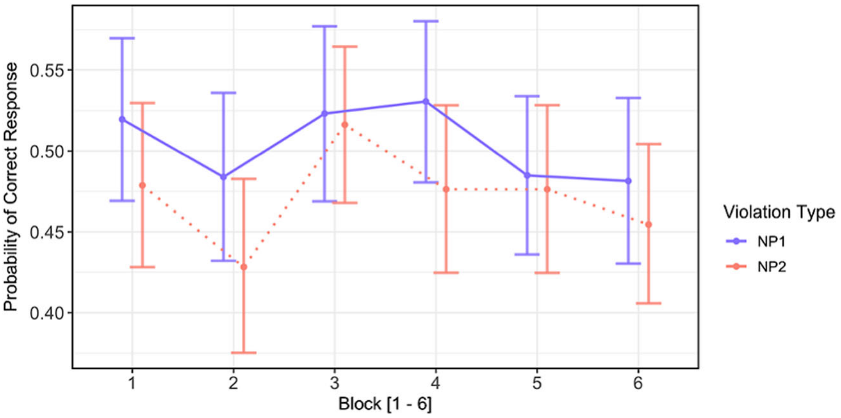

Fig. 11 Comparison in performance between classifier-noun violations at the position of the first (NP1) and second noun (NP2) phrases. Probability of correct response is (y-axis) modelled across block (x-axis; 1–6) for NP1 (purple solid line) and NP2 violations (orange dotted line). Bars represent the 83% confidence interval

Indefrey, & Hagoort, 2016). These studies have shown that humans are able to rapidly acquire complex grammatical rules, demonstrating above chance-level performance within initial exposure (Folia et al., 2010) and high performance (e.g., > 80% accuracy) after prolonged learning (i.e., 9 days; Weber et al., 2016). Using Mini Pinyin, the current study demonstrated that monolingual native English speakers can attain a moderate degree of accuracy on a judgement task after one learning session; however, this was dependent on specific word order rules. For AVU, UVA and A*bă*UV sentences, participants were able to obtain up to 100% accuracy, while the detection of classifier-noun violations, on average, remained at chance level. Such variability in performance is higher than in previous higher-order language learning experiments (Friederici et al., 2002; Mueller et al., 2005), possibly due to the greater number of word order rules and vocabulary. Further, classifiers are a linguistic property not present in English, and as such, participants may have found it difficult

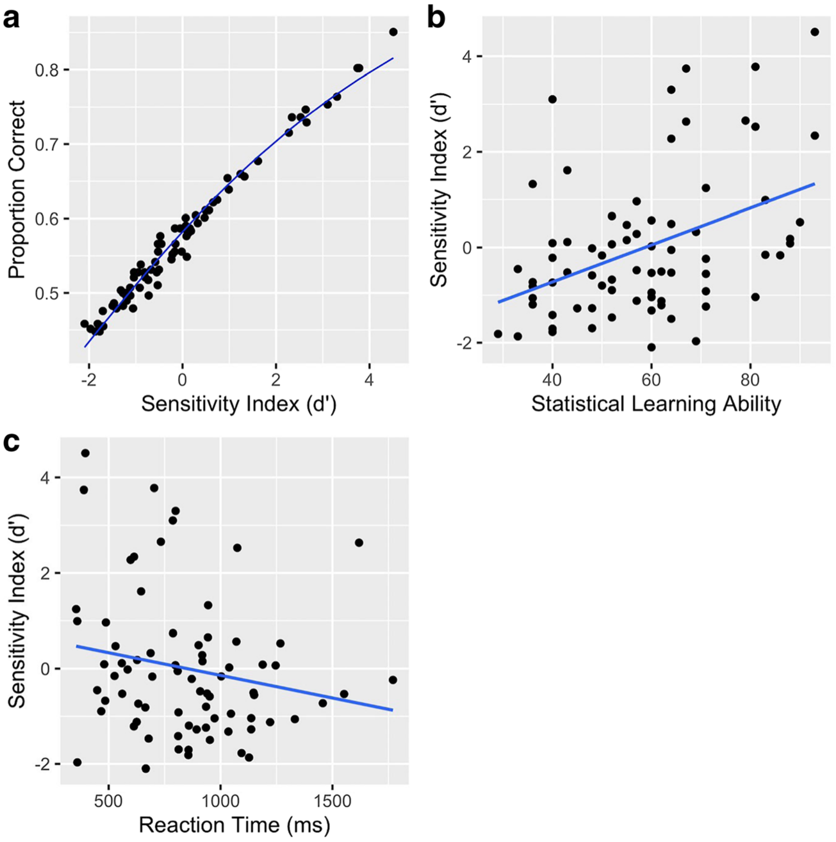

Fig. 12 Scatterplots illustrating the relationship between (**a**) proportion correct and d’, (**b**) statistical learning ability and d’ and (**c**) reaction time (ms) and d’

to encode the dependencies between classifier-noun pairs. From a complementary learning systems perspective (Davis & Gaskell, 2009; Norman, 2010), prior knowledge is critical for successful encoding and retention of new (linguistic) knowledge. When there is a dissociation between mnemonic information and existing schemata, the time for adequate consolidation is longer and may require greater offline reprocessing (e.g., during sleep; Cross et al., 2018; Zion, Nevat, Prior, Bitan, 2019). Consequently, our sample of native English speakers may have preferentially learned word order regularities over classifier rules, given that word order is a prominent cue for sentence interpretation in English (Bornkessel-Schlesewsky et al., 2011; MacWhinney et al., 1984). Following this line of interpretation, we might have observed greater consolidation (reflected in higher performance) for classifier violations if participants had a greater consolidation opportunity, particularly if it also included a period of sleep, given the role of sleep in memory consolidation (Diekelmann & Born, 2010; Rasch & Born, 2013). Another possible, and perhaps less complex, interpretation of the difficulty in learning classifier-noun dependencies relates to their salience. During the sentence learning task, there was no direct visual input corresponding to classifier-noun relations, and thus, classifiers may have required more time to be learned regardless of prior knowledge. In order to test these hypotheses, future research should include a control group of participants whose native language contains classifier-like rules, such as in Japanese (Mitsugi, 2018; Sudo, 2016), and compare differences in learning to participants who have no prior linguistic experience with classifier-noun rules.

Comparison between Mini Pinyin and existing paradigms

It is important to note the key distinctions between previous paradigms and the current MML. First, Mini Pinyin contains a number of word order permutations, including subject- and object-initial and verb-medial and verb-final constructions, compared to existing MML paradigms that contain only verb-final constructions (e.g., Mueller, 2006). In particular, the use of ambiguous word orders (i.e., UVA constructions), which requires interpretation via animacy cues (human > animal > inanimate object), necessitates the learning of word order rules that deviate from subject-initial constructs relying primarily on linear order-based predictions for interpretation, such as those present in English. Further, in the verb-final constructions, role assignments to the arguments are rendered unambiguous via the coverbs *bạ̌* and *bèi*, and when combined with Undergoer-initial constructions (i.e., *bèi* constructions), allow for the study of how more complex word order rules are learned in short time periods. Mini Pinyin also contains 47 vocabulary items, including 16 verbs and 25 nouns, compared to the relatively small vocabulary in previous paradigms (e.g., 17; Mueller, 2006). Unlike Mini-Nihongo, the lexicon of Mini Pinyin also contains nouns that differ in animacy, allowing for the clarification of whether learners use animacy to establish thematic structure in sentences with fixed word orders, such as U*bèi*AV constructions (Lamers, 2006).

Another important consideration is that we did not use a learning criterion during the sentence learning and judgement phases, unlike in other studies (e.g., Friederici et al., 2002; Mueller, Hirotani, & Friederici, 2007), where participants needed to attain a high degree of proficiency before completing further testing. Furthermore, Mini-Nihongo has almost exclusively been studied in the auditory modality, while Mini Pinyin is currently designed in the visual domain. This is an important distinction, given that auditory presentation can provide additional predictive cues (e.g., prosodic information; Snedeker & Trueswell, 2003) regarding upcoming words, while rapid serial visual presentation – as was used here – inherently differs from auditory language (Kyriaki, Schlesewsky, Bornkessel-Schlesewsky, 2020) as well from as natural reading (cf. Kretzschmar, Bornkessel-Schlesewsky, & Schlesewsky, 2009; Rayner & Clifton, 2009). For example, the duration of word presentation, as well as the interval between each word, may influence learning by increasing working memory demands (Busler & Lazarte, 2017; de Liaño, Potter, & Rodríguez, 2014). This difference in learning criteria and modality (visual versus auditory) between the current study and previous paradigms makes it difficult to compare behavioural performance between Mini Pinyin and previous AGL/MML studies. As such, future research using Mini Pinyin may wish to compare differences across the learning and judgement tasks on trained and untrained participants and between modalities, particularly when using neuroscientific measurements, which may reveal subtle differences in neural activity between groups (trained, untrained) and conditions (visual, auditory). However, despite the above-mentioned differences with previous paradigms, we have demonstrated that native English speakers are able to learn a comparatively large vocabulary and complex grammatical rules in a relatively short time period. Future research may wish to determine whether native speakers of languages (e.g., German and Turkish) that rely more heavily on other cues (e.g., animacy) learn Mini Pinyin at a different rate to native English speakers.

How can Mini Pinyin be applied to characterise higher-order language learning?

Of the studies examining higher-order language learning, Kepinska et al. (2017a, 2017b) have shown that learning success may depend on individual language learning aptitudes. From a more domain-general perspective, however, (adult) language learning may be at least, in part, dependent on individual differences in statistical learning ability. Indeed, evidence from behavioural studies has demonstrated that individual differences in human statistical learning predicts variations in grammatical processing abilities (Misyak & Christiansen, 2012; Misyak, Christiansen, & Tomblin, 2010; Kidd, Donelley, & Christiansen, 2018; Siegelman, Bogaerts, Christiansen, & Frost, 2017b). In the current study, we demonstrated that statistical learning ability significantly predicted learning of Mini Pinyin, explaining approximately 20% of the variance in overall performance on the sentence judgement task. However, in Mini Pinyin, there are a number of dependencies between individual words: classifiers are paired with specific noun phrases, and the inclusion of the coverbs *bǎ* and *bèi* generates verb-final constructions, as well as indicating different Actor/Undergoer word orders. Here, we also demonstrated that statistical learning ability predicted the probability of a correct response for verb-position violations but had no association with Actor/Undergoer order violations. This suggests that statistical learning does not account for role assignments based on relational semantic information (e.g., animate/inanimate nouns). As such, while it might capture aspects of a typical AGL paradigm (i.e., the order of category sequences), it does not predict individual aptitude for the relational processing of nonadjacent (semantic) information.

From this perspective, higher statistical learning ability appears to facilitate the extraction of sequence-based grammatical rules, resulting in more efficient encoding and generalisation of the statistical cues necessary for fixed word order sentence interpretation. However, it is important to note that we did not experimentally manipulate statistical learning ability or exposure to Mini Pinyin. Therefore, despite the strong relationship between statistical learning ability and language learning, and evidence beyond this study, we cannot establish causation. Further, statistical learning ability has been shown to correlate with verbal working memory (Misyak & Christiansen, 2012), which has also been shown to strongly predict individual differences in linguistic processing abilities (Archibald, 2016). From this perspective, other domain-general mechanisms, such as working memory, may explain language learning ability. Future research should directly manipulate statistical learning ability and/or other parameters related to language learning, including working memory capacity. This may involve varying exposure time to Mini Pinyin during the learning phase and determining whether this influences performance on the judgement task.

Mini Pinyin can also be used to study the role of animacy-related information in higher-order language learning. In the present study, participants were required to learn undergoer-initial word orders (i.e., UVA), and thus likely predicted that the second noun phrase was the Actor, given that all verbs were transitive. Further, the distinction between grammatical and ungrammatical *bǎ* and *bèi* sentences via NP order critically hinges on animacy. By manipulating the animacy status of noun phrases in *bǎ* and *bèi* constructions and in undergoer-initial word orders, such as in UVA sentences, future research could characterise how animacy-related cues facilitate the extraction of sentential rules, and whether this is modulated by domain-general mechanisms, including statistical learning ability, verbal working memory and/or general intelligence.

Finally, we argue that Mini Pinyin could be used to characterise the neurobiological mechanisms underlying cross-linguistic differences in sentence processing. In native sentence processing, there are qualitative differences in ERP responses across languages (Bornkessel-Schlesewsky et al., 2011). For example, semantic reversal anomalies (syntactically well-formed but semantically implausible sentences; e.g., *the apple ate the boy*) elicit an N400 effect in comparison to control sentences (e.g., *the boy ate the apple*) in German but not in English (Bornkessel-Schlesewsky et al., 2011; cf. Kyriaki, Schlesewsky, & Bornkessel-Schlesewsky, 2020). Given that Mini Pinyin contains word orders where the first noun phrase is inanimate, future studies could track the point at which learners demonstrate native-like language-related ERP components during the processing of non-canonical word orders. More recent work has also studied the role of neural oscillations in sentence processing (for review: Meyer, 2018), revealing that the predictability of upcoming words manifests in distinct patterns of oscillatory activity and accurate sentence comprehension (Molinaro, Monsalve, & Lizarazu, 2016). For example, an increase in beta oscillatory power (~ 13–30 Hz) is argued to reflect accurate predictions of upcoming words based on the sentential context (Bastiaansen & Hagoort, 2015; Lam, Schoffelen, Udden, Hulten, & Hagoort, 2016; Lewis & Bastiaansen, 2015; Wang et al., 2012). From this perspective, Mini Pinyin could be used to test the role of neural oscillations during language learning and sentence processing. For example, oscillatory activity could be recorded at the position of verb violations in fixed word order sentences (i.e., AbaUV) to better characterise the role of neural oscillations in linear order-based sentence processing. Given that we provide all materials – including vocabulary items, sentences and picture stimuli, experimental tasks and statistical analysis scripts – future researchers can easily adapt their experimental parameters to address these outstanding questions.

Additional methodological considerations

Early work (Friederici et al., 2002; Kepinska, de Rover, Caspers, & Schiller, 2016; Mueller et al., 2007; cf. Weber et al., 2016) has employed traditional ANOVA analyses on aggregated data, removing by-participant and by-item variance. Such approaches also tend to convert continuous data (e.g., language proficiency) into factorial categories (e.g., high vs. low language proficiency; Kepinska, de Rover, et al., 2017b; Kepinska, Pereda, et al., 2017), which can result in a reduction in variance and statistical power (Cunnings, 2012; Link & Cunnings, 2015; Quenè & Bergh, 2008). Here, we utilised generalised linear mixed model analyses with by-participant and by-item random effects, likely increasing the power and generalisability of the observed effects (Alday, 2019; Baayen, Davidson & Baates, 2008). For example, some participants may have had a steeper learning curve than others, and some sentence items may have been more difficult to process than others. The presence of by-item and by-participant random effect structures thus allows interpretations on outcome measures to be based more strongly on experimental manipulations and not the influence of uncontrolled variance, which can increase the risk of committing a type I error (Meteyard & Davies, 2020).

Despite the strengths of the current study, there are several limitations that need to be acknowledged. First, the current form of Mini Pinyin is in the visual domain, with sentences presented via rapid serial visual presentation. Rapid serial visual presentation differs from natural auditory language and naturalistic reading (which contains the opportunity to regress to previous regions of a sentence; Kretzschmar, Bornkessel-Schlesewsky, & Schlesewsky, 2009; Rayner & Clifton, 2009), and as such, may influence the mechanisms underlying the extraction and generalisation of grammatical rules. Second, our sample consisted of monolingual native English speakers. While this was done to control the influence of linguistic experience, speakers of other languages, like German and Turkish, which share similar properties to Mandarin Chinese (e.g., morphological case marking may be akin to Mandarin coverbs), may learn the regularities of Mini Pinyin at a different rate to English speakers.

Another factor to consider is the potential influence of the orthography and phonology of Mini Pinyin on learning. Several studies (e.g., Akamatsu, 2003; Hamada & Kodo, 2008; Wang, Koda, & Perfetti, 2003) have shown that orthographic knowledge of one's native language influences word learning in a second language. In particular, English is an alphabetic language, while Chinese is a morphographic language, and as such, processing relies more heavily on phonetic and visual information, respectively (Miller, 2018). To address these differences in orthography and phonology, each word in Mini Pinyin was presented in an English alphabet- and script-based (i.e., Roman) system. Further, each word was pronounceable in English, such that they were legal pseudowords (as listed in Table 3, which summarises the vocabulary of Mini Pinyin), and while there may have been different degrees of difficulty in the mapping of graphemes to phonemes, words were counterbalanced between sentence conditions. From this perspective, any potential influence of orthography on learning could not account for between-condition differences in grammaticality judgements. Moreover, we assume that potential effects of individual word difficulty will have been captured by the by-item variability in the mixed-effect models (Baayen et al., 2008; Quené, Huub, & Bergh, 2008).

It is also important to consider whether Mini Pinyin can be used to study language learning in highly multilingual populations, particularly where knowledge of Mandarin Chinese (or typologically similar languages) is prevalent. Here, we were interested in studying the learning trajectory of Mini Pinyin while controlling for prior linguistic knowledge. As such, it was critical that we excluded individuals who had previous exposure to Mandarin Chinese. However, similar to studies using Mini-Nihongo (e.g., Mueller, 2006; Mueller et al., 2005, 2007), future research may be interested in comparing individuals with and without knowledge of the language within which the paradigm is based. From this perspective, future research will need to carefully consider the use of Mini Pinyin in relation to their research question and the populations of prospective participants they have access to.

Finally, we tested participants' knowledge immediately after learning. As discussed above, the generalisation of grammatical regularities may require longer consolidation periods. Indeed, the consolidation of language-related rules has been shown to occur up to 36 hours after learning (Zion et al., 2019). From this perspective, testing participants' knowledge of Mini Pinyin at varying time points after learning, particularly after a period of sleep, may help to characterise both the rate of learning and the mechanisms underlying the generalisation of complex grammatical rules.

Conclusions

We have demonstrated that native monolingual English speakers are able to learn a complex miniature language containing various word order permutations after a short exposure period. We also showed that individual differences in statistical learning ability is positively associated with performance on the judgement task and that this effect is modulated by prior (linguistic) knowledge. The linguistic properties of Mini Pinyin, including fixed and flexible word orders, classifiers and coverbs, builds upon existing artificial grammar paradigms and offers a novel method for characterising the neurobiological mechanisms underlying cross-linguistic differences in sentence processing. By providing all stimulus and task materials, future studies will be able to better characterise the psycholinguistic, neurophysiological and neuroanatomical correlates of higher-order language learning and incremental sentence processing.

Acknowledgements We thank Andrew Corcoran and Louise Kyriaki for valuable feedback on an earlier version of this manuscript. We also thank our participants for their enthusiastic engagement throughout this study.

Funding Preparation of this work was supported by Australian Commonwealth Government funding awarded to ZC under the Research Training Program (RTP; number 212190). IB-S is supported by an Australian Research Council Future Fellowship (FT160100437). LZ-W is supported by an Australian Government RTP scholarship.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests.

Open practices statement All data and materials are available at https://osf.io/f5ksg/?view_only=738200b63e78433eb50e667ea772c08f. This experiment was not preregistered.

References

Akamatsu, N. (2003). The effects of first language orthographic features on second language reading in text. *Language Learning*, *53*(2), 207–231.

Alday, P. M. (2019). How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting itslimits. *Psychophysiology, 56*(12), e13451. https://doi.org/10.1111/psyp.13451.

Archibald, L. M. (2016). Working memory and language learning: A review. *Child Language Teaching and Therapy*, *33*(1), 5–17. https://doi.org/10.1177/0265659016654206

Austin, P. C., & Hux, J. E. (2002). A brief note on overlapping confidence intervals. *Journal of Vascular Surgery, 36*(1), 194–195. https://doi.org/10.1067/mva.2002.125015

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.

Bastiaansen, M., & Hagoort, P. (2015). Frequency-based segregation of syntactic and semantic unification during online sentence level languagecomprehension. *Journal of Cognitive Neuroscience, 27*(11), 2095–2107. https://doi.org/10.1162/jocn_a_00829.

Bates, D. M. (2010). lme4: Mixed-effects modeling with R.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Bolker, M. B. (2015). Package 'lme4'. *Convergence*, *12*(1), 2.

Bates, E., Devescovi, A., & Wulfeck, B. (2001). Psycholinguistics: A Cross-Language Perspective. *Annual Review of Psychology*, *52*(1), 27.

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In *The Crosslinguistic Study of Sentence Processing* (pp. 3-76). New York: Cambridge University Press.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. *Cognition, 11*(3), 245–299.

Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*(4), 787–821. https://doi.org/10.1037/0033-295X.113.4.787

Bornkessel-Schlesewsky, I., Kretzschmar, F., Tune, S., Wang, L., Genc, S., Philipp, M., … Schlesewsky, M. (2011). Think globally: Cross-linguistic variation in electrophysiological activity during sentence comprehension. *Brain and Language*, *117*(3), 133–152. https://doi.org/10.1016/j.bandl.2010.09.010

Bornkessel-Schlesewsky, I., Schlesewsky, M., Small, S. L., & Rauschecker, J. P. (2015). Neurobiological roots of language in primate audition: Common computational properties. *Trends in Cognitive Sciences*, *19*(3), 142–150. https://doi.org/10.1016/j.tics.2014.12.008

Busler, J. N., Lazarte, A. A., (2017). Reading time allocation strategies and working memory using rapid serial visual presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1375-1386. https://doi.org/10.1037/xlm0000392

Cribari-Neto, F., & Zeileis, A. (2009). Beta regression in R.

Cross, Z. R., Kohler, M. J., Schlesewsky, M., Gaskell, M. G., & Bornkessel-Schlesewsky, I. (2018). Sleep-dependent memory consolidation and incremental sentence comprehension: computational dependencies during language learning as revealed by neuronal oscillations. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00018

Cross, Z. R., Randolph F. Helfrich, R F., Kohler, M. J., Corcoran, A. W., Coussens, S., Zou-Williams, L., Schlesewsky, M. M., Gaskell, M. G., Knight, R. T., Bornkessel-Schlesewsky, I. (2020). Slow wave-spindle coupling during sleep predicts language learning and associated oscillatory activity. *bioRxiv*.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369-382.

Daltrozzo, J., Emerson, S. N., Deocampo, J., Singh, S., Freggens, M., Branum-Martin, L., & Conway, C. M. (2017). Visual statistical learning is related to natural language ability in adults: An ERP study. *Brain and Language*, *166*, 40–51. https://doi.org/10.1016/j.bandl.2016.12.005

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

de Diego-Balaguer, R., Fuentemilla, L., & Rodriguez-Fornells, A. (2010). Brain Dynamics Sustaining Rapid Rule Extraction from Speech. *Journal of Cognitive Neuroscience*, *23*(10), 3105–3120. Retrieved from pbh. (95462200)

de Liaño, B. G., Potter, M. C., & Rodríguez, C. (2014). Working memory effects in speeded RSVP tasks. *Psychological Research*, *78*, 124–135. https://doi.org/10.1007/s00426-013-0479-7

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews. Neuroscience*, *11*(2), 114–126. https://doi.org/10.1038/nrn2762

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. https://doi.org/10.1016/j.tics.2015.07.013

Dorrian, J., Lamond, N., & Dawson, D. (2000). The ability to self-monitor performance when fatigued. *Journal of Sleep Research*, *9*(2), 137–144.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108.

Folia, V., Uddén, J., De Vries, M., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. *Language Learning*, *60*, 188–220.

Fox, J. (2011). *Tests for Multivariate Linear Models with the car Package*. 99.

Franzen, P. L., Siegle, G. J., & Buysse, D. J. (2008). Relationships between affect, vigilance, and sleepiness following sleep deprivation. *Journal of Sleep Research*, *17*(1), 34–41.

Friederici, A. D. (2005). Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends in Cognitive Sciences*, *9*(10), 481–488. https://doi.org/10.1016/j.tics.2005.08.008

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(1), 529–534. https://doi.org/10.1073/pnas.012611199

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of Chinese individual classifiers. *Language & Cognitive Processes*, *24*(7–8), 1124–1179. https://doi.org/10.1080/01690960802018323

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. *Trends in Cognitive Sciences* https://doi.org/10.1016/j.tics.2017.04.013

Hamada, M., & Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning*, *58*(1), 1–31.

Hayes, N. A., & Broadbent, D. E. (1988). Two modes of learning for interactive tasks. *Cognition*, *28*(3), 249–276.

Her, O. S., Chen, Y. C., & Yen, N. S. (2017). Neural correlates of quantity processing of Chinese numeral classifiers. *Brain and Language*, *176*, 11–18. https://doi.org/10.1016/j.bandl.2017.10.007

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of Sleepiness: A New Approach. *Psychophysiology*, *10*(4), 431–436. https://doi.org/10.1111/j.1469-8986.1973.tb00801.x

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

Hopper, P., & Traugott, E. (2003). Grammaticalization. Cambridge: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jost, E., & Christiansen, M. H. (2017). Statistical learning as a domain-general mechanism of entrenchment.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54.

Kepinska, O., de Rover, M., Caspers, J., & Schiller, N. O. (2017a). Connectivity of the hippocampus and Broca's area during acquisition of a novel grammar. *Neuroimage*, *165*, 1–10. https://doi.org/10.1016/j.neuroimage.2017.09.058

Kepinska, O., de Rover, M., Caspers, J., & Schiller, N. O. (2017b). Whole-brain functional connectivity during acquisition of novel grammar: Distinct functional networks depend on language learning abilities. *Behavioural Brain Research*, *320*, 333–346. https://doi.org/10.1016/j.bbr.2016.12.015

Kepinska, O., Pereda, E., Caspers, J., & Schiller, N. O. (2017). Neural oscillatory mechanisms during novel grammar learning underlying language analytical abilities. *Brain and Language*, *175*, 99–110. https://doi.org/10.1016/j.bandl.2017.10.003

Kepinska, O., de Rover, M., Caspers, J., & Schiller, N. O. (2016). On neural correlates of individual differences in novel grammar learning: An fMRI study. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2016.06.014

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

Kretzschmar, F., Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). Parafoveal vs. Foveal N400s dissociate spreading activation from contextual fit. *NeuroReport*, *20*, 1613–1618.

Kyriaki, L., Schlesewsky, M., & Bornkessel-Schlesewsky., I. (2020). Semantic reversal anomalies under the microscope: Task and modality influences onlanguage-associated event-related potentials. *European Journal of Neuroscience*. https://publons.com/publon/10.1111/ejn.14862.

Lam, N. H., Schoffelen, J. M., Uddén, J., Hultén, A., & Hagoort, P. (2016). Neural activity during sentence processing as reflected in theta, alpha, beta, andgamma oscillations. *Neuroimage, 142*, 43–54. https://doi.org/10.1016/j.neuroimage.2016.03.007.

Lamers, M. J. (2006). Cracking the nutshell differently. Commentary on Müller. *Language Learning, 56,* 271–277.

Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.8. https://CRAN.R-project.org/package=emmeans. 1.4.8. https://CRAN.R-project.org/package=emmeans

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex,68*, 155–168. https://doi.org/10.1016/j.cortex.2015.02.014.

Li, P., Bates, E., & MacWhinney, B. (1993). Processing a language without inflections: A reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language*, *32*(2), 169.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185–207.

MacGregor-Fors, I., & Payton, M. E. (2013). Contrasting diversity values: statistical inferences based on overlapping confidence intervals. *PLoS One*, 8(2).

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 127–150. https://doi.org/10.1016/S0022-5371(84)90093-8

Mathot, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal ofNeuroscience, 48*(7), 2609–2621. https://doi.org/10.1111/ejn.13748.

Miller, R. T. (2018). English orthography and reading. *The TESOL Encyclopedia of English Language Teaching*. https://doi.org/10.1002/9781118784235.eelt0461.

Milne, A. E., Mueller, J. L., Mannel, C., Attaheri, A., Friederici, A. D., & Petkov, C. I. (2016). Evolutionary origins of non-adjacent sequence processing in primate brain potentials. *Scientific Reports*, *6*, 36259. https://doi.org/10.1038/srep36259

Mirkovic, J., & Gaskell, M. G. (2016). Does Sleep Improve Your Grammar? Preferential Consolidation of Arbitrary Components of New Linguistic Knowledge. *PLoS One*, *11*(4), e0152489. https://doi.org/10.1371/journal.pone.0152489

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302–331.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *1*, 31.

Mitsugi, S. (2018). Generating predictions based on semantic categories in a second language: A case of numeral classifiers in Japanese. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2017-0118

Molinaro, N., Monsalve, I. F., & Lizarazu, M. (2016). Is there a common oscillatory brain mechanism for producing and predicting language?. *Language,Cognition and Neuroscience, 31*(1), 145–158. https://doi.org/10.1080/23273798.2015.1077978.

Mueller, J. L., Bahlmann, J., & Friederici, A. D. (2010). Learnability of embedded syntactic structures depends on prosodic cues. *Cognitive science, 34*(2),338–349. https://doi.org/10.1111/j.1551-6709.2009.01093.x.

Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, *17*(8), 1229–1244.

Mueller, J. L. (2006). L2 in a Nutshell: The investigation of second language processing in the miniature language model. *Language Learning*, *56*, 235–270. https://doi.org/10.1111/j.1467-9922.2006.00363.x

Mueller, J. L., Hirotani, M., & Friederici, A. D. (2007). ERP evidence for different strategies in the processing of case markers in native speakers and non-native learners. *BMC Neuroscience*, *8*, 18. https://doi.org/10.1186/1471-2202-8-18

Mueller, J. L., Milne, A., & Männel, C. (2018). Non-adjacent auditory sequence learning across development and primate species. *Current Opinion in Behavioral Sciences*, *21*, 112–119.

Mueller, J. L., Rueschemeyer, S. A., Ono, K., Sugiura, M., Sadato, N., & Nakamura, A. (2014). Neural networks involved in learning lexical-semantic and syntactic information in a second language. *Frontiers in Psychology*, *5*, 1209. https://doi.org/10.3389/fpsyg.2014.01209

Narula, S. C. (1979). Orthogonal polynomial regression. *International Statistical Review/Revue Internationale de Statistique*, 31–36. https://doi.org/10.2307/1403204

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus*, *20*(11), 1217–1227. https://doi.org/10.1002/hipo.20855

Opitz, B., & Friederici, A. D. (2007). Neural basis of processing sequential and hierarchical syntactic structures. *Human Brain Mapping*, *28*(7), 585–592. https://doi.org/10.1002/hbm.20287

Opitz, Bertram, & Friederici, A. D. (2003). Interactions of the hippocampal system and the prefrontal cortex in learning language-like rules. *NeuroImage*, *19*(4), 1730–1737. https://doi.org/10.1016/s1053-8119(03)00170-8

Petersson, K. M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, *120*(2), 83–95. https://doi.org/10.1016/j.bandl.2010.08.003

Poletiek, F. H., & Lai, J. (2012). How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: a statistical account. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 367*(1598), 2046–2054. https://doi.org/10.1098/rstb.2012.0100

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.

R Core Team (2020). R: *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews*, *93*(2), 681–766. https://doi.org/10.1152/physrev.00032.2012

Rayner, K., & Clifton, C. J. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, *80*(1), 4–9.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 88.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 492.

Roehr-Brackin, K., & Tellier, A. (2019). The role of language-analytic ability in children's instructed second language learning. *Studies in Second Language Acquisition*, *41*(5), 1111–1131.

Rohrmeier, M. A., & Cross, I. (2014). Modelling unsupervised online-learning of artificial grammars: Linking implicit and statistical learning. *Consciousness and Cognition*, *27*, 155–167. https://doi.org/10.1016/j.concog.2014.03.011

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *Journal of Memory and Language*, *110*, 104038.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017b). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *372*(1711), 20160059.

Siegelman, N., Bogaerts, L., & Frost, R. (2017a). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods* https://doi.org/10.3758/s13428-016-0719-z

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language, 48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory and measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 18.

Sudo, Y. (2016). The semantic role of classifiers in Japanese. *Baltic International Yearbook of Cognition, Logic and Communication*, *11*(1).

Uddén, J., Ingvar, M., Hagoort, P., & Petersson, K. M. (2012). Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: Investigating the push-down stack model. *Cognitive Science*, *36*, 1078-1101. https://doi.org/10.1111/j.1551-6709.2012.01235.x

Uddén, J., Ingvar, M., Hagoort, P., & Petersson, K. M. (2017). Broca's region: A casual role in implicit processing of grammars with crossed non-adjacent dependencies. *Cognition*, *164*, 188-198. https://doi.org/10.1111/j.1551-6709.2012.01235.x

Uddén, J., & Männel, C. (2018). Artificial grammar learning and its neurobiology in relation to language processing and development. In *The Oxford Handbook of Psycholinguistics* (pp. 755–783). Oxford University Press.

Van Dongen, H. P., Olofsen, E., Dinges, D. F., & Maislin, G. (2004). Mixed-model regression analysis and dealing with interindividual differences. In *Methods in enzymology* (Vol. 384, pp. 139–171). Academic Press.

Wang, L., Schlesewsky, M., Bickel, B., & Bornkessel-Schlesewsky, I. (2009). Exploring the nature of the 'subject'-preference: Evidence from the online comprehension of simple sentences in Mandarin Chinese. *Language & Cognitive Processes*, *24*(7–8), 1180–1226.

Wang, L., Schlesewsky, M., Philipp, M., & Bornkessel-Schlesewsky, I. (2012). The role of animacy in online argument interpretation in Mandarin Chinese. In *Case, word order and prominence* (pp. 91–119). Springer, Dordrecht.

Wang, M., Koda, K., & Perfetti, C. A. (2003). Alphabetic and non-alphabetic L1 effects in English semantic processing: A comparison of Korean and Chinese English L2 learners. Cognition, 87, 129–49.

Weber, K., Christiansen, M. H., Petersson, K. M., Indefrey, P., & Hagoort, P. (2016). FMRI Syntactic and Lexical Repetition Effects Reveal the Initial Stages of Learning a New Language. *The Journal* *of Neuroscience*, *36*(26), 6872–6880. https://doi.org/10.1523/JNEUROSCI.3180-15.2016

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., & Petkov, C. I. (2013). Auditory artificial grammar learning in macaque and marmoset monkeys. *The Journal of Neuroscience*, *33*(48), 18825–18835. https://doi.org/10.1523/JNEUROSCI.2414-13.2013

Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, *16*, 16.

Zion, D. B., Nevat, M., Prior, A., & Bitan, T. (2019). Prior knowledge predicts early consolidation in second language learning. *Frontiers in Psychology,10*, 2312. https://doi.org/10.3389/fpsyg.2019.02312.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.