Laboratory Phonology
Journal of the Association for Laboratory Phonology

Tang, K., & Baer-Henney, D. (2023). Modelling L1 and the artificial language during artificial language learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), pp. 1–54. DOI: https://doi.org/10.16995/labphon.6460

Open Library of Humanities

Modelling L1 and the artificial language during artificial language learning

Kevin Tang*, Department of English Language and Linguistics, Institute of English and American Studies, Heinrich-Heine-University, Düsseldorf, 40225, Germany; Department of Linguistics, University of Florida, Gainesville, Florida, 32611-5454, United States of America, kevin.tang@hhu.de

Dinah Baer-Henney, Department of Linguistics, Heinrich-Heine-University, Düsseldorf, 40225, Germany, dinah.baer-henney@hhu.de

*****Corresponding author.

Artificial language learning research has become a popular tool to investigate universal mechanisms in language learning. However, often it is unclear whether the found effects are due to learning, or due to artefacts of the native language or the artificial language, and whether findings in only one language will generalise to speakers of other languages. The present study offers a new approach to model the influence of both the L1 and the target artificial language on language learning. The idea is to control for linguistic factors of the artificial and the native language by incorporating measures of wordlikeness into the statistical analysis as covariates. To demonstrate the approach, we extend Linzen and Gallagher (2017)'s study on consonant identity pattern to evaluate whether speakers of German and Mandarin rapidly learn the pattern when influences of L1 and the artificial language are accounted for by incorporating measures assessed by analogical and discriminative learning models over the L1 and artificial lexicon. Results show that nonwords are more likely to be accepted as grammatical if they are more similar to the trained artificial lexicon and more different from the L1 and, crucially, the identity effect is still present. The proposed approach is helpful for designing cross-linguistic studies.

1. Introduction

A central topic in linguistics is seeking to understand how humans learn linguistic patterns in order to become competent speakers. Universal and language-specific mechanisms are said to guide the learner through the learning progress (e.g., Baer-Henney, Kügler, & van deVijver, 2015; Culbertson, Smolensky, & Legendre, 2012; Kakolu Ramarao, Tang, & Baer-Henney, 2023; Moreton, 2008). To investigate these mechanisms, a powerful tool which is gaining increasing popularity is the artificial language learning (ALL) paradigm. The paradigm allows for monitoring countless aspects of language learning. With a relatively simple laboratory setting, we can observe mechanisms guiding language acquisition, language learning, and comparisons thereof. In such an experiment, an artificial miniature lexicon is governed by a certain grammatical pattern. Participants are first exposed to the artificial language and in a subsequent perception or production test it is examined whether the pattern has been learned.

Artificial languages are not only useful to track the acquisition path of the language-learning child (Berko, 1958; Kakolu Ramarao, Zinova, Tang, & van de Vijver, 2022; van de Vijver & Baer-Henney, 2014); artificial languages have also been used to investigate learning mechanisms in both children (Chambers, Onishi, & Fisher, 2003; Cristià & Seidl, 2008; Culbertson & Newport, 2015; White & Sundara, 2014) and adults (Baer-Henney & van de Vijver, 2012; Carpenter, 2010; Finley, 2011, 2017; Finley & Badecker, 2009; Martin & White, 2021; Moreton, 2008; Onishi, Chambers, & Fisher, 2002; Pater & Tessier, 2003; Wilson, 2006). Researchers have uncovered evidence for learning mechanisms, the so-called *biases* that are said to guide the learner during learning process (e.g., Baer-Henney et al., 2015; Carpenter, 2010; Culbertson & Newport, 2015; Culbertson et al., 2012; Finley, 2011, 2017; Finley & Badecker, 2009; Martin & White, 2021; Moreton, 2008; Tang, Kakolu Ramarao, & Baer-Henney, 2022; van de Vijver & Baer-Henney, 2014; Wilson, 2006). The consequence of biases is that certain phonological patterns are learned with more ease than others. Linzen and Gallagher (2017), for instance, have shown that English speaking participants rapidly take up a phonological pattern from an artificial language that requires two consonants of the artificial stimulus to be identical. After short training the identical consonant pattern is predominantly generalised (as compared to a non-identical consonant pattern). Learning mechanisms like this one are viewed as underlying learning mechanisms that all humans share.

However, when utilising the ALL paradigm, there is the risk that the learning effects experimenters observe actually are the unintended byproducts of the native language of the learner and/or the artificial language to be learned during the experiment. The present paper offers a possibility to disentangle a possible learning effect from artefacts of the native and the artificial language during artificial language learning. In order to provide convincing evidence, ALL research should fulfill the following criteria. First, convincing evidence for the universality of mechanisms should, in fact, address more languages or find replicated effects in multiple

languages. Second, a possible effect should not be able to be explained by characteristics of the native language of the speaker nor by characteristics of the artificial language used in the paradigm but only due to learning. Fulfilling both criteria can pose some methodological problems as we will discuss in the following section. In the present paper, we offer an approach to address these problems. The approach facilitates designing cross-linguistic studies by statistically controlling for linguistic factors of the artificial and the native language. Several measures of wordlikeness are incorporated into the analysis as covariates.

1.1. Challenges in stimulus design in ALL research

Stimuli in ALL studies are typically designed by combining a subset of the L1 phoneme inventory (Baer-Henney et al., 2015; Skoruppa, 2019; White et al., 2018), or by simply replacing sounds in subsyllabic positions of real words (Keuleers & Brysbaert, 2010). This is to ensure that the items conform to the phonotactics of the participants' L1. We believe that the desire to come up with AL stimuli that are similar and yet different to the L1 bears some risks to overlook peculiarities of the L1 that possibly have an impact on the outcome of an ALL study. In the remainder of this section, we discuss the possible challenges ALL research faces and a number of problems that arise from this methodology.

When conducting an ALL study researchers face several challenges. One of the challenges concerns the scenario of investigating the same mechanism with different speaker populations. Commonly, the same set of language materials is created for all different language groups in order to make a better comparison. However, the language material used cannot contain linguistic properties that vary too widely from the native languages. As the number of languages involved in the study increases, the design space of the material becomes more restricted. Let us consider a cross-linguistic study such as White et al. (2018)'s, which could raise some difficulties. The most basic criterion in this approach is to ensure the languages share the same phonemes. See **Figure 1** for an illustration of the problem. In the best scenario, the number of overlapping phonemes would be reasonably large if the two languages are within the same language family (e.g., English and German) (**Figure 1b**). If the two languages are more distinct (e.g., German and Mandarin), the number of overlapping phonemes is small (**Figure 1a**). However, to better establish the universality of learning mechanisms, one must increase the number of languages as well as selecting languages that are more distinct, resulting in a highly reduced subset of phonemes (**Figure 1c** and **Figure 1d**).

With only a small subset of phonemes available, the research potential of the ALL paradigm becomes limited in terms of the type of mechanisms that can be examined and the number of artificial language stimuli used in the experiment (henceforth: nonwords) that can be generated. In sum, ALL experiments in phonology usually ensure that the miniature artificial language broadly conforms to the typical word shapes of the learners' L1 in order to minimise the potential

negative effects of being too different from the learners' L1. For instance, learners could fail to perceive the nonwords correctly due to novel phonotactic patterns. At the same time, the nonwords in the miniature artificial language are designed to be different from the real words of the learners' L1 in order to minimise the interference from existing word knowledge. This balance of trying to be similar and yet different from the participants' L1 is difficult to strike. This is particularly an issue for cross-linguistic studies if a mechanism is found to be at work only in some language groups but not others. This could be driven by a difference in how this balance was struck across the language groups.

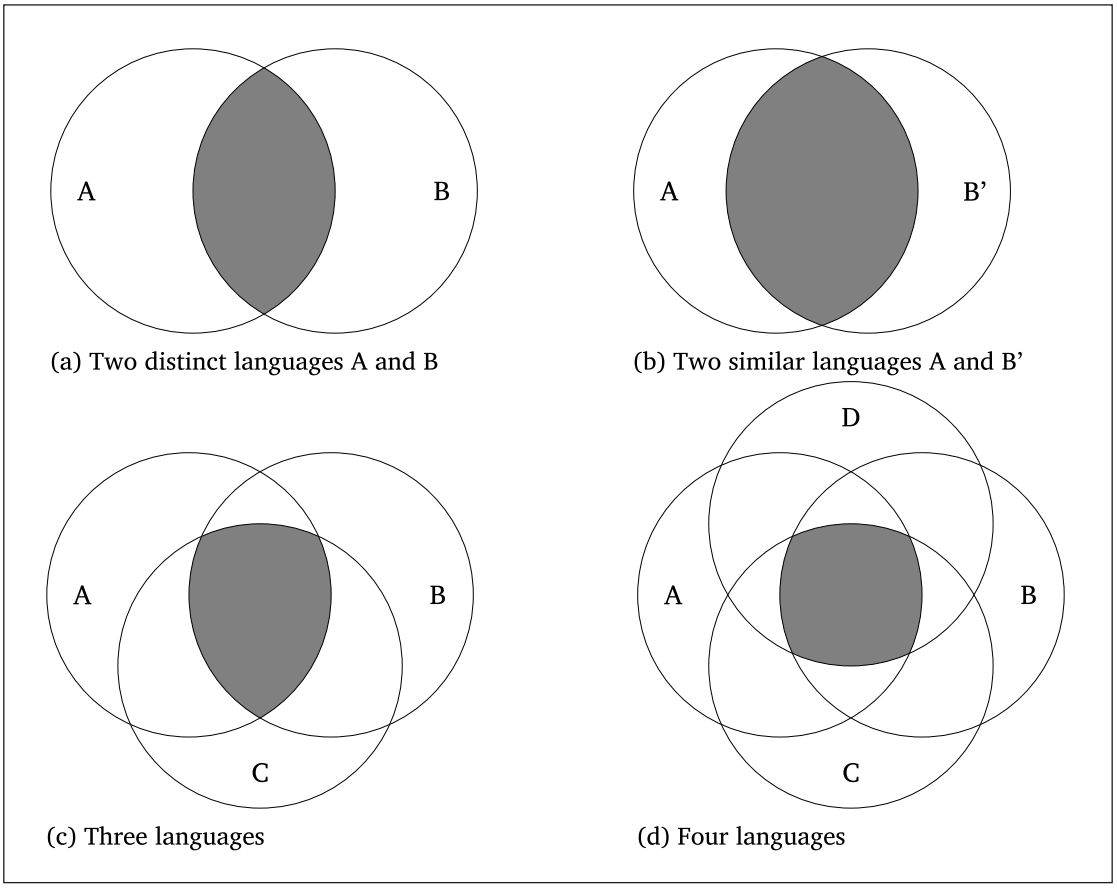

Figure 1: Decreasing phoneme overlap (gray area) as more languages (circles) are taken into account.

A second challenge is to cope with the fact that different native languages can influence the learning differently. A wealth of literature on Second Language Acquisition (see e.g., Edwards & Zampini, 2010; Kroll & De Groot, 2009, for an overview) has highlighted the role of language transfer in second language learning. When learning a new language, L1 influences how L2 (and L3) is learned. It has been shown that the same L2 is learned differently by learners with different

L1s (Iverson & Evans, 2007, 2009; Schepens, van Hout, & Jaeger, 2020), and the similarity between L1 and L2 has been shown to have both a facilitatory effect and an inhibitory effect on language learning. Now ALL experiments on phonological research questions typically overlook the possible interaction of their L1 and the AL. Learners are faced with an artificial language, but the AL is not controlled for possible L1 effects. A particular L1 might interact differently with different artificial languages (see Onnis and Thiessen (2013) for such an effect in syntactic learning). The specific choice of speaker population to be tested can have a consequence for the outcome of the study.

One risk is that researchers could select languages that work *against* the mechanism of interest to best evaluate its robustness. For instance, an ALL study by Yin and White (2018) found an effect of homophony avoidance with L1-English participants. In this study, L1-English learners were biased against phonological patterns that create homophony. However, the fact that English does not have a high degree of homophony could be driving the effect of homophony avoidance. Therefore, this study could benefit from a replication with speakers of a homophone-rich language, such as Mandarin Chinese, to evaluate the universality of homophony avoidance. Consider the following two hypothetical examples: (1) In the specific choice of artificial items chosen by the researchers, they could overlook hidden "subpatterns" in L1. A specific nonword might resemble 80% of L1 words that follow the pattern of interest while the L1 as a whole does not display a pattern. For instance, in an experiment teaching a fronting pattern experimenters might overlook how the specific phonological contexts in the nonwords are accidentally very similar to a subpart of the learners' L1 lexicon in which front sounds are very common, even though the overall lexicon does not display such a rule. (2) Researchers could over-rely on broad generalisations (e.g., L1 has vowel harmony or L1 is a trochaic language). The information that vowel harmony may only apply to a certain word class or that trochees are predominantly found in disyllabic words but not in trisyllabic words may be missed.

Let us take a closer look at what solutions to the abovementioned problems the ALL and psycholinguistic literature offers, and which problems are still to be solved. One approach used by *some* studies is to run a control condition in which there is only the test phase without training (Finley & Badecker, 2012), or a control condition with an uninformative training phase not containing critical items (Finley, 2011). The control group is then compared with the experimental group that received training. There are two main issues with the control group approach of dealing with L1 effects. First, the strength of L1 might be different across the two groups. Let us assume that the L1 has a similar effect as the hypothesised learning mechanism on the AL test items. The control group approach relies on the assumption that the L1 affects across the control group and the experimental group equally because the difference in the responses of the test items is interpreted as evidence for the learning mechanism at work. However, should the amount of L1 affect the two groups *unequally*, then the effect of the mechanism could be masked

or falsely supported by the difference in the responses of the groups. The effect might be masked when the L1 effect is stronger in the control group than in the experimental group because, for instance, there is simply no AL input to rely on. Subsequently, the difference between the two groups are evened out. The effect might be falsely supported when the L1 effect is stronger in the experimental group than in the control group because, for example, the statistical patterns of the AL input primes those same patterns in the L1 thus enhancing the L1 effect. In other words, languages which already display a systematic pattern (in either direction) in their L1, and therefore in a control group, will not be evaluated accordingly, since there would be no possibilities to disentangle the influence of the mechanism at work and prior knowledge. Second, the control group approach could lead to potential misinterpretation of seemingly contradicting findings from different language groups. There is the risk that differences between control and experimental groups are bigger in one language population but smaller or absent in another language population. This would, in fact, suggest that the mechanism is not universal.

Another approach is to construct AL stimuli that have been controlled for their L1 wordlikeness. The following studies examine the existence of the patterns of interest using a lexicon. Baer-Henney (2015), for instance, used frequency counts of types and tokens to establish hypotheses regarding the outcome of the experiment if participants were to mirror lexical statistics. Similarly, Seidl and Buckley (2005) excluded a pre-existing bias on the basis of a global check of the lexicon in which they find that both patterns under investigation occur equally often. Other studies dedicated parts of their discussion to a simple check of lexical frequencies and find their results matching (Greenwood, 2016) or mismatching (Finley & Badecker, 2012; Myers & Padgett, 2014) the lexical statistics, concluding whether lexical knowledge could have explanatory power in the specific case or not. What these studies have in common is that they check and/or discuss lexical statistics, but they do not incorporate them into the analysis. These studies evaluate lexical statistics that concern the pattern of interest in the artificial language and they suggest not testing a population of speakers in which the pattern is already part of the language. While this way a pattern of interest may be under control (e.g., a pattern related to consonants), another one may not (e.g., a potential pattern related to vowels). It is thus possible that the researcher unintentionally uses particularly wordlike nonwords or that certain nonwords are similar to a subpart of the L1 lexicon.

In fact, in psycholinguistic literature outside the context of ALL, efforts in evaluating the wordlikeness of nonwords across multiple languages have been similarly limited with a few exceptions. For instance, in nonword repetition research, Boerma et al. (2015) and Howell et al. (2017) have created quasi-universal nonwords for the purpose of designing a nonword repetition task that would function well with a large number of languages with minimal bias towards a particular language. In these studies the nonwords are controlled for their phonological structures and the number of phonological neighbours. In phonological research, T.-Y. Chen

and Myers (2021) created a platform for collecting wordlikeness judgements in an effort to encourage researchers to share their judgement data towards a compilation of a mega database of cross-linguistic wordlikeness judgements. While researchers could select nonwords that are similarly wordlike with this database, this project is still a work in progress and the languages involved are very limited. Methodologically, there are nonword generation tools that can generate nonwords based on user-specified lexical properties (e.g., Duyck, Desmet, Verbeke, & Brysbaert, 2004; Keuleers & Brysbaert, 2010). However, they were not designed for selecting cross-linguistic nonwords (at least not without further development; see Eden (2018) for a cross-linguistic approach of measuring phonological distances). Finally, this approach falls short of its extendibility to multiple languages especially if the languages are typologically different, since the more languages we try to take into account, the harder it would be to create nonwords that are similarly wordlike across the languages. In sum, we believe that this approach is not appropriate for ALL research on the universality of learning mechanisms and in the following section we introduce an approach to improve the control over the design of an artificial language in ALL research.

1.2. A lesson from psycholinguistics

Already in 1996, a critical review of the ALL paradigm by Redington and Chater (1996, p. 129) highlighted the importance of item control. It was stated that "Just as psycholinguists routinely control for word frequency, cloze value, and the like, so researchers should routinely control for factors such as bigram and trigram frequencies, if they intend to eliminate hypotheses based on knowledge of such simple fragments, rather than, for example, memory for whole strings, the extraction of an underlying grammar, and so forth."

Building on this insight, the present paper harnesses methodologies in psycholinguistic literature in lexical processing and human learning: Rather than controlling unwanted L1 and AL factors in the experimental design through stimuli selection, we take lexical statistics of L1 and AL into account in a regression model (Baayen 2004, 2010). Controlling for L1 and AL effects means that we monitor whether wordlikeness of test items to L1 words or to the artificial training items played a role in learning. Subjects could prefer test words that are similar to the L1 lexicon and to the AL lexicon in two ways. A test item can be similar to the L1/AL and can be supported by L1/AL lexical statistics in terms of a) the target AL pattern of interest, and b) other non-target patterns. In fact, the methodological decisions made in an ALL study by Boll-Avetisyan and Kager (2016) exemplifies the necessity of including lexical statistics in a regression model since controlling unwanted factors through stimuli selection is not always possible. In their study, the authors attempted to control for wordlikeness of L1. During stimuli selection, the authors were able to ensure that their nonwords were balanced across conditions in term of their positional syllable frequencies, transitional probabilities between phonemes, and lexical neighbourhood

density. However, they were unable to further control for cohort density, a measure related to lexical neighbourhood density. As a result, the authors added cohort density as a covariate, which turned out to a significant variable, in the statistical models in order to capture its potential L1 influence on the learners when tested with the AL materials.

This approach serves two particular purposes in ALL. First, in an ALL experiment with only one language group, the effects of L1 and AL can be regressed to reveal the existence of the learning effect, even without a control group. Second, in an ALL experiment with multiple language groups, the effects of the respective L1s can be regressed to evaluate the universality of the learning effect. This approach overcomes the aforementioned shortcomings of using a control group and using stimuli selection to control for L1 effects, offers a control for the interference of the AL lexicon, as well as gives additional insights in language learning. Regressing L1 and AL effects would give us the ability to understand L1 and AL effects and in how they jointly affect learning. Learners may learn other patterns of the AL lexicon, as well as the intended patterns that are relevant to the learning effect of interest. By controlling wordlikeness of test items to training items with regression, we can better understand whether the learning effect is at play. We are therefore less likely to miss the effect of an effect even if the effect was inherently small and possibly masked by L1 and AL. If this were the case, no difference between control and experimental groups would arise. By taking L1 and AL into account as covariates, subtle effects have a better chance of being detected. By regressing effects of multiple L1s instead of a categorical variable of language groups, researchers would be able to understand the cause of any potential language group differences. The statistical model enables us to better separate the native language effect from the learning effect by having gradient language factors (compared to a categorical factor of language groups). An example of this approach to understanding cross-linguistic differences can be found in a phonetic study by Burchfield and Bradlow (2014) on the syllable reduction effect in Mandarin and English (see Günther, Smolka, and Marelli (2019) for another example). The authors observed cross-linguistic differences, with Mandarin showing more syllable reduction than English. Crucially, this language group difference was reducible to the differences in the relative proportions of open and closed syllables across these languages. In other words, the observed cross-linguistic differences in the dependent variable are determined by specific cross-linguistic structural differences.

We illustrate our proposed approach by evaluating the impact of the L1 in artificial language learning. Linzen and Gallagher (2017)'s study on the identity effect in English will serve as a starting point for the investigation of the relevance of several factors from the L1 and AL lexicons. We adopted their paradigm because it investigates a general type of learning (not as specific as dealing with a learning bias), and the findings will be applicable to all learning studies, including the strand of ALL studies concerning possibly universal biases. Linzen and Gallagher (2017) trained English speakers with the same number of artificial words containing

identical and non-identical consonants. Natural language is likely to have fewer words containing identical consonants than those containing non-identical consonants (Graff & Jaeger, 2009; see references within Tang & Akkuş, 2022). Thus, the training set exhibited an overrepresentation of consonant identity nonwords relative to chance and this led speakers to be more likely to accept new items as part of the learned language if they conformed to the identity pattern. We describe the details of the experiment in the section below. Linzen and Gallagher (2017) did run a control version of their experiment to rule out the possibility of a pre-existing bias resulting from words of the L1 as compared to resulting from exposure phase. Participants from the control group received no training and showed no preference for identical consonant frames while participants from the experimental group showed a preference. Nonetheless, it was acknowledged by the authors that the possible influence of L1 and the memory of AL exposure material seems to be an understudied problem in ALL research (Linzen & Gallagher, 2017, p.26).

The paper is organised as follows: Section 2 outlines the details of our study. We first introduce the logic of the original study by Linzen and Gallagher (2017) before we present the materials (2.1) of the two artificial languages for our Mandarin and German language groups, followed by the methodological details of our study (2.2). Section 2.3 introduces the variables of interests and their implementations. In this section, we will introduce the language-dependent variables (Section 2.3.1) and language-independent variables (Section 2.3.2). Section 3 presents our results for German and Mandarin speakers. Section 4 reviews our findings regarding the original identity effect, task effects, and L1 and AL effects before closing with a section on implications for learning research. Section 5 concludes the paper.

2 The present study

We adopted Linzen and Gallagher (2017)'s ALL experiment in which the authors detected learning of an identity pattern after short exposure.

Linzen and Gallagher (2017) tested English speakers. In their experiment participants listened to artificial language input during training. During the test, they were asked if unattested and attested items could be accepted as part of the artificial language. Specifically, we adapted their version 2a of the experiment (Linzen & Gallagher, 2017, p.10), in which the pattern to be learned is a probabilistic abstract generalization and does not make reference to the phonetic properties of any particular sound: We adjusted stimuli and training length to our needs. The artificial language input in training was ambiguous in the sense that half of the items conformed to an identity constraint; the other half did not. Items conforming to the identity constraint contained an identical consonant in each consonant position. We refer to these C_aVC_aV items as items with identical C_C_ frames. Items not conforming to the identity constraint contained different consonants in each consonant position. We refer to these C_aVC_bV items as items with non-identical C_C_ frames. During the test, participants were asked if attested and unattested

items (with the same number of identical and non-identical C_C_ frames) could be accepted as part of the artificial language. Note that in this design there is not a correct or an incorrect answer to whether a given stimulus belongs to the artificial language exposed to the participants since there were equal number of identical and non-identical C_C_ frames during the training and test phase.[1] Despite equal input of identical and non-identical C_C_ frames, participants were more likely to accept the identity pattern as compared to a non-identical items. This captures the preference with which participants are willing to take up patterns and generalise them to new material. While the input contains more identical nonwords than non-identical nonwords relative to chance, participants show rapid learning of this phonotactic pattern, namely the overrepresentation of identical nonwords. This tendency to overestimate the identity pattern as part of the new language is what Linzen and Gallagher (2017) refer to as identity effect.

The present study aims at extending the examination to two radically distinct languages and uses the method as a test case to investigate the role of the lexicon of both the native language and the AL in ALL. For the present study we created a Mandarin and a German artificial language version. Using a language-adapted AL lexicon we familiarised German and Mandarin speakers with an artificial language and aimed to (1) replicate the original effect in two other languages, and (2) also track if and how L1 and AL lexical knowledge contributes to our learners' behaviour. We investigated several types of lexical knowledge that could have influenced performance in an ALL task.

To measure the influence of a learner's L1 (German/Mandarin) and AL (the training nonwords) on the task, we selected three lexical variables that reflect different types of lexical knowledge. We investigated the influence of (1) activation diversity, a measure of distributions of co-occurring features across the lexicon during lexical selection, (2) neighbourhood density, a classic lexical variable for modelling analogical learning, and (3) the probability of the specific C_C_ frame, a variable which highlights a potential attention effect by the participants to the CVCV template as it was used by all of the nonwords in the experiment. While the first two variables control the artificial languages' similarity to the learners' L1 in a general way, the last variable, consonant frame identity, controls more directly for AL similarity. Consonant frame probability is a control variable that looks at the specific pattern in the artificial item and tells us about how often its specific pattern is found in the L1 lexicon. Neighbourhood density (the measure resulting from an Generalized Neighbourhood Model (GNM)) and activation diversity (the Naive Discriminative Learning (NDL) measure) look at L1 similarity more generally. Together they enable us to examine the role of the lexicon at different levels of specificity – over any general dimensions (vowels, consonants, or features) or over the target AL pattern.

[1] Attested items consisted of attested consonant frames but different vowels, hence no single test item resurfaced in test after training. While one could argue that accepting a test item with an attested frame is a correct answer, this categorisation of correctness does not influence how we analyse the data.

In sum, the present study is a cross-linguistic study on learning an identity constraint while controlling for language-specific variables and as such it addresses several questions of interest and is applicable to other ALL studies as well. First, we aim to see whether the effect under investigation (here: identity) is a robust effect even if we take language-specific variables into account. By including L1 variables we test whether the identity effect originates from or may be masked by from the speakers' L1. By including AL variables we test whether the identity effect originates from or may be masked by the design of the artificial language used in the experimental learning situation. Only if the effect under investigation turns out to be robust, even when we take into account language-specific effects, we can argue for its existence. Second, if the effect under investigation turns out to show up consistently across language populations we are able to argue for its universality.

In the following sections we describe the materials (Section 2.1) and the technical details of the method (Section 2.2), and finally we report how we prepared for the evaluation of language-dependent and language-independent variables under investigation (Section 2.3).

2.1. Materials

We created a Mandarin and German artificial miniature lexicon, out of which we chose individual training and test items for the ALL experiment. All items had a CVCV syllable structure with identical or non-identical consonant frames and tense language-specific vowels. The lexicon contained exposure items as well as attested and unattested test items. Half of these items conform to an identity constraint consonants and the other half does not.

Linzen and Gallagher (2017) tracked the learning of their artificial language with different amounts of exposure. The effect became apparent after only little exposure with few tokens per exposure item type. The identity effect was found in attested items after using two exposure tokens per type and in unattested items after using four exposure tokens per type (Linzen & Gallagher, 2017, experiment 2a). We therefore decided to use an equivalent set with three exposure sets to investigate influences of the native as well as the artificial lexicon.

Participants were assigned to one of four experimental groups – there was a German as well as a Mandarin version, and a frequent and an infrequent condition per language. For each group we sampled experimental stimuli from a group-specific item pool. In what follows we explain how stimuli for these pools were created before we turn to the question how stimuli were distributed during the experiment.

2.1.1 Stimuli creation

For the Mandarin and German version of the experiment we used a subset of language-specific consonants, since selecting a set of overlapping consonants of two typologically distinct

languages would severely limit the number of possible nonwords, a problem that was illustrated in **Figure 1**. Our aim was to prepare four miniature lexicons: Two per language, one where preselected consonant frames were relatively frequent and one where consonant frames were relatively infrequent according to the lexical statistics of the native language (L1). Token frequency was used as a proxy for L1 wordlikeness to select consonant frames for each subgroup[2]. The frequency of consonant frames was estimated by calculating their Zipf values (a normalised frequency scale: log10 of the frequency per million plus three) (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) in language-specific tokens.[3] Each participant was presented with either the more frequent frames or the less infrequent frames of one language.[4]

The consonants used to construct the consonant frames were counterbalanced across two conditions: Identity (identical vs. non-identical frames) and attestedness (attested vs. unattested frames). Concretely, two non-overlapping sets of consonants were used to construct the consonant frames, with one set being used to construct unattested, identical frames and attested, non-identical frames (e.g., [m, p^h, h, k^h] in high frequent German condition) and another set being used to construct attested, identical frames and unattested, non-identical frames (e.g., [b, t^h, n, ʁ] in high frequent German condition). See **Table 1**, which summarises consonant frames used in the two languages, German and Mandarin, together with their frequency status. For half of the participants we set up a corresponding item set in which consonant frames of the exposure items/attested test items and unattested test items were interchanged.

Equipped with language-specific consonant frames we generated all possible CVCV sequences using the Mandarin vowel set { aː, ai, au, əi, iː, eː, oː, uː, ʏː, əu, yː } and the German vowel set {aː, eː, iː, oː, uː}.[5] Four constraints were applied to come up with a final set of items. First, existing words were excluded. Second, we avoided token similarity by ensuring the first vowel of the items was not identical (we avoided stimuli such as naːno, naːne). Third, we avoided token similarity by ensuring the second vowel of the items was not identical (we avoided stimuli such as naːno, neːno). Fourth, we avoided vowel metathesis patterns across items (we avoided stimuli such as naːno, noːna). In the few remaining cases when we had to pick the violated forms, we picked the one that did not share a vowel at a certain position with the test item. If we had still no choice,

[2] In principle, type frequency and other frequency-related lexical estimates can also be used. We encourage the readers to examine other estimates in their individual studies.

[3] Consonant frames not found in our lexicons are smoothed to have a Zipf value of 1.

[4] This high-low frequency condition is highly correlated with the consonant frame probability variable and, in fact, they are the most correlated pair of variables, with $R = 0.3$ for German and $R = 0.6$ for Mandarin. Furthermore, the high-low frequency condition is less fine-grained than the continuous consonant frame probability variable. Therefore, this high-low frequency condition was not considered as a fixed effect in the statistical models.

[5] For both languages, tense vowels were chosen to best match the L&G's study, which also chose tense vowels that are either long vowels or diphthongs. The number of nonwords that can be generated for Mandarin using only long vowels were too small to match the number of German nonwords; therefore, diphthongs were also chosen for Mandarin.

we chose the one where the second vowel was identical to the test item. Of the remaining items we generated their syllable-bigram[6] probability and picked the four CVCV tokens evenly across the spectrum of syllable-bigram probability. Mandarin is a tonal language with four lexical tones which are the high tone (tone 1), the rising tone (tone 2), the low tone (tone 3), and the falling tone (tone 4). Mandarin items were realised with the falling tone (tone 4) on both syllables. The tone sequence 4–4 is the most common disyllabic word forms in Mandarin (Lin 2016). German items were trochaic. A stressed penultimate before an unstressed ultimate corresponds to the most common word forms in German (Wiese, 2000). An example item set for one German participant[7] is illustrated in **Table 2**; a full list of artificial items for both German and Mandarin speakers in all groups can be found in Tables 10, 11, 12, and 13 of the appendix.

The two miniature item sets were recorded separately – one by a native speaker of Mandarin (Henan dialect) and one by a native speaker of Standard German. Items were recorded in the Mandarin carrier sentence "请你把 __ 再说一遍" (Pinyin: qing3 ni3 ba3 __ zai4 shuo1 yi1 bian4.) (*en: Please say __ once again.*) or the German carrier sentence "Ich habe noch nie __ gehört." (*en: I never heard __ before.*). The stimuli were recorded in an anechoic booth in the phonetics laboratory at Heinrich-Heine University Düsseldorf. We extracted stimuli and scaled their intensity to 70 dB using Praat (Boersma & Weenink, 2018).

2.1.2. Stimuli distribution

For each participant, there was an individual training and test set consisting of stimuli from the experimental group-specific item pool (frequent German, infrequent German, frequent Mandarin, infrequent Mandarin). For each of the language-specific eight attested consonant types (four identical and four non-identical), each participant was given three randomly selected tokens per type (out of four tokens) as the training set, and the remaining fourth token was used as an attested test item. For each of the eight unattested consonant types (four identical and four non-identical), each participant was given one randomly selected token per type (out of four tokens) as part of the unattested test items. In total, each participant was given 24 tokens (three tokens x eight attested consonant types) in the training set, eight tokens (one token x eight attested consonant types) in the attested test set, and eight tokens (one token x eight unattested consonant types) in the unattested test set. Presentation order of exposure and test items was randomised within the experimental phase.

[6] A syllable bigram consists of two consecutive syllables.

[7] The table shows a simplified version of items not using diacritics for reasons of readability.

			Exposure/Test (attested)		Test (unattested)	
			Frame	**Token$_{Zipf}$**	**Frame**	**Token$_{Zipf}$**
German	High frequency	Identical	b_b_	5.506	m_m_	5.294
			t^h_t^h_	5.047	p^h_p^h_	5.219
			n_n_	4.500	h_h_	4.283
			ʁ_ʁ_	4.015	k^h_k^h_	4.085
		Non-identical	m_k^h_	3.731	ʁ_t^h_	5.158
			h_p^h_	3.519	b_ʁ_	5.127
			p^h_m_	3.072	t^h_n_	3.984
			k^h_h_	1	n_b_	2.072
	Low frequency	Identical	g_g_	3.644	v_v_	2.373
			l_l_	3.563	z_z_	1.595
			t͡s_t͡s_	1.896	j_j_	1
			f_f_	1	d_d_	1
		Non-identical	z_d_	3.887	l_g_	5.274
			d_v_	2.709	t͡s_l_	4.943
			v_j_	1	f_t͡s_	1
			j_z_	1	g_f_	1
Mandarin	High frequency	Identical	t_t_	5.711	t^h_t^h_	5.391
			k^w_k^w_	4.784	x^w_x^w_	4.809
			w_w_	4.559	l_l_	4.517
			s_s_	4.055	t͡s^w_t͡s^w_	4.234
		Non-identical	x^w_l_	5.829	t_w_	4.755
			l_t͡s^w_	3.181	w_k^w_	4.286
			t^h_x^w_	3.107	s_t_	3.783
			t͡s^w_t^h_	2.319	k^w_s_	3.190
	Low frequency	Identical	k^{hw}_k^{hw}_	3.976	p^h_p^h_	4.052
			t͡s^{hw}_t͡s^{hw}_	3.087	t͡ʂ^{hw}_t͡ʂ^{hw}_	3.743
			p^j_p^j_	2.905	n^j_n^j	3.280
			ʐ_ʐ_	2.753	t^{hw}_ t^{hw}_	3.030
		Non-identical	t^{hw}_n^j_	4.306	k^{hw}_ʐ_	2.6784
			p^h_t^{hw}_	3.456	p^j _t͡s^{hw}_	2.1733
			n^j_t͡ʂ^{hw}_	1	t͡s^{hw}_k^{hw}_	1
			t͡ʂ^{hw}_p^h_	1	ʐ_p^j_	1

Table 1: Consonant frames and their lexical distributions in L1 lexicons.

	Training			Test	
				Attested	Unattested
Identical C_C	be:bo	ba:bu	bo:bi	bu:ba	kʰo:kʰi
	tʰi:tʰa	tʰe:tʰu	tʰo:tʰi	tʰa:tʰo	ha:hu
	na:ni	no:nu	nu:ne	ne:no	pʰe:pʰa
	ʁe:ʁa	ʁi:ʁo	ʁu:ʁi	ʁa:ʁu	mi:mu
Non-identical C_C	kʰo:ha	kʰe:hi	kʰi:hu	kːʰu:ho	bo:ʁu
	ha:pʰi	hi:pʰu	hu:pʰo	ho:pʰe	ʁi:tʰu
	pʰa:mi	pʰu:me	pʰo:ma	pʰe:mu	tʰe:ni
	mo:kʰa	mu:kʰe	ma:kʰi	me:kʰo	no:ba

Table 2: Illustration of one item set presented to a participant in the frequent German group.

2.2. Methods

Experiments were run online using Experigen (Becker & Levine, 2013). Participants were asked to wear headphones. During training, every training item was played once. During the test, participants were asked if testing items could be accepted as words from the new language. The instructions used by Linzen and Gallagher (2017) were translated with minor variations into German and Mandarin Chinese and can be found in Appendix A. After the experiment, participants were asked to fill out a short demographic background sheet asking for age, gender, and language history.

Participants were recruited via social media platforms and mailing lists. They provided informed consent and they were offered the chance to enter a raffle to win a gift voucher. Two hundred and thirty-two native German adults took part in the German experiment. Their mean age was 31.5 years, ranged between 18 and 70. One hundred and thirty-six were women, 50 were men, two identified as other. Forty-four participants did not provide information about their gender. In the Mandarin version of the experiment, 219 Mandarin native speakers took part, with a mean age of 21.5, ranging from 15–55. Ninety seven were women, 95 were men, and 27 did not provide gender information.

2.3. Variables under investigation

For the language-dependent variables, we computed them with respect to their L1 and AL. For the nonwords in each version of the experiment, the L1 variable was computed over the full L1 lexicon (German or Mandarin), and the AL variable was computed over the corresponding artificial language that was exposed to the participants. Thus, the AL variables were computed over the individual artificial exposure lexicon which was much smaller than the L1 lexicon.

To better estimate the L1 lexical knowledge of our participants, we chose frequency lists compiled using subtitle texts, namely SUBTLEX-DE (Brysbaert et al., 2011) and SUBTLEX-CH (Cai & Brysbaert, 2010) for German and Mandarin respectively. This was motivated by the fact that lexical frequencies derived from subtitle texts have consistently shown to outperform those from other genres in capturing behavioural responses in psycholinguistic tasks across languages (Brysbaert & New, 2009; de Chene, 2014; Keuleers, Brysbaert, & New, 2010; Tang, 2012; Tang & de Chene, 2014; Tang & Shaw, 2021). These reference lexicons were then enriched with IPA. The German lexicon was created by combining the lexical entries in the German section of CELEX (Baayen, Piepenbrock, & Gulikers, 1995) with SUBTLEX-DE's frequency estimates. The Mandarin lexicon was created by transcribing the lexical entries in SUBTLEX-CH with CEDICT (Denisowski, 1997) and the pinyin pronunciation guide outlined in Duanmu (2007).[8] Surprasegmental information was removed for German but not for Mandarin because German stress has a low lexical functional load (Surendran & Niyogi, 2003; Tang, Chang, et al., 2022) and is predictable (Féry, 1998) compared to Mandarin tones. ALL studies have consistently found that learners are sensitive to phonological features in the training stimuli; therefore, we compute our wordlikeness variables on the featural level as opposed to on the segmental level (Durvasula & Liter, 2020; Finley, 2022; Linzen & Gallagher, 2017) In addition, we also investigated a number of language-independent factors such as trial number, reaction time, and the original finding of the acceptability rate of identical patterns (Linzen & Gallagher, 2017). We controlled for whether C_C_ frames were attested or unattested during training.

2.3.1. Language-dependent variables

Three language-dependent predictors were considered and are described in detail in the next paragraphs. As measures of wordlikeness, we used activation diversity and neighborhood density, two variables that are relatively uncorrelated (Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017). Third, we considered consonant frame probability.

Activation diversity Assessing activation diversity of the nonword test stimuli used is one way to account for wordlikeness of our stimuli. In this way we are able to control for how similar the stimuli are to the words of the native lexicon of the speaker, and we are able to control for how similar the stimuli are to the nonwords of the artificial training lexicon of the speaker. The following section describes how the measures are assessed. Activation diversity is a measure resulting from an Naive Discriminative Learning (NDL) model. NDL incorporates the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972).

[8] We opted to use the surface representation for both German and Mandarin. This is motivated, in part, by how the underlying representations of Mandarin Chinese are very theory-dependent and cannot be separated into syllable parts very easily (Duanmu, 2007).

In this framework the learning process consists of *cues* (the set of input units) and *outcomes* (the set of output units). What is a cue and what is an outcome is specified in the specific learning scenario. Cues and outcomes can either be present or be absent. The learning rule captures the change in association strength between cues and outcomes. Presence and absence of cues and outcomes determine strength of associations in the network and – as a consequence – learnability of the outcomes. If a cue is not present in a learning event, then no change in association strength is made. If a cue and an outcome are both present, then the association strength increases. If a cue is present but an outcome is not present, then the association strength decreases. If a cue or an outcome has not been encountered before, then the association strength does not change. To put simply, the strength of association between a cue and an outcome depends on whether the cue is predictive of the outcome. Cues compete for predictive values based on whether a cue successfully predicts an outcome or not. Please see Appendix B for the mathematical details of the Rescola-Wagner equations (Rescorla & Wagner, 1972).

The Rescorla-Wagner learning rule has been shown to provide a psychologically plausible model of human learning in a number of lexical and phonological processing tasks. In lexical decision tasks, NDL-derived variables have been found to outperform classical lexical-distributional measures (Milin et al., 2017). NDL was able to model morphological variations in Russian and provide complementary information to other modelling approaches (logistic regression and decision trees and random forest) (Baayen, Endresen, Janda, Makarova, & Nesset, 2013). Crucially, it has also been applied in ALL studies to understand the mechanism underlying language learning (Vujović, Ramscar, & Wonnacott, 2021). Recent studies have shown that latent variables derived from a cue-to-outcome matrix can capture behavioural responses in a number of psycholinguistic tasks, such as lexical decision and word naming, similarly well compared to a number of classic psycholinguistic variables, such as token frequency, family size, and phonotactic probability.[9]

In lexical processing, typically an NDL model is trained on certain types of sublexical cues (e.g., bigrams of letters or acoustic characteristics), (input units) and words as outcomes (output units) (see Milin et al., 2017; Nixon, 2020). It is designed to discriminate between words on the basis of sublexical cues. The model results in a cue-to-outcome matrix and each cell in the matrix contains the weight of a specific cue activating a specific outcome. As the model is being exposed to each word and its corresponding cues, the activation weights are updated accordingly. In this way, the NDL model represents the mental lexicon, which is dynamic and flexible. The matrix then offers the possibility to check the wordlikeness of new bigram combinations. Thus, activation values can be derived, stored as L1 and AL activation diversity measures, and used as an additional predictor in our final model.

[9] For examples of comparisons between NDL-derived latent variables and traditional psycholinguistic variables, please review Hendrix (2016) for a series of studies on word naming, compound reading, and picture naming tasks in English, and Pham and Baayen (2015) for similar comparisons in a lexical decision task in Vietnamese.

Thus, the latent variable that is of our particular interest is the activation diversity given a set of cues. In the cue-to-outcome matrix, each cue is a vector of the same length as the total number of outcomes (nonwords for the AL exposure lexicon, and words for the L1 lexicon).

Unlike previous studies such as Milin et al. (2017), the cues of our NDL models were not computed over bigram letters but rather over bigram phonological features. The feature specification of each phone was taken from the PHOIBLE feature chart (Moran, McCloy, & Wright, 2014). Following the framework of autosegmental phonology (Goldsmith, 1976), the bigrams were computed separately over segment tiers and tonal tiers. While segments were broken down into features, tones were not, and the tone values were used to compute tonal bigrams directly. Concretely, this means that in our assessment, we used featural bigram representations such as a bigram of features [+syllabic] followed by [-syllabic] as cues to check against the outcomes.

For each test nonword we assessed the activation diversity for L1 and AL. The ndl2 library (version 0.1.0.9002) was used to implement our NDL models (Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011; Shaoul et al., 2015). With respect to the L1, we first trained an NDL model using an L1 lexicon to obtain a cue-to-outcome matrix, where cues are bigram features and the outcomes were real words of German or of Mandarin. To then obtain the L1 activation diversity value for each nonword stimulus in the testing phase, we converted each nonword into a featural bigram representation and looked up the activation values that correspond to each featural bigram cue using the cue-to-outcome matrix.[10] The sum of the absolute values of the activation values for all the cues of a nonword is its activation diversity value. With respect to the AL activation diversity, the only difference is the lexicon that the NDL model was trained on first. Instead of a German/Mandarin lexicon, an NDL model was trained using the individual miniature AL lexicon which contains only the exposure items per participant. Thus, for each test nonword we computed an L1 activation diversity value and an AL activation diversity value. In this way, we interpret activation diversity as a measure of wordlikeness of a nonword. A test nonword that is very wordlike with respect to the reference lexicon (L1 or AL) would have cues with high activation values in the cue-to-outcome matrix and thus would have a high (L1 or AL) activation diversity value.

To illustrate the NDL procedure to gain activation diversity values, a miniature example is provided in **Table 3**. **Table 3** shows that how L1 activation diversity values for the two test

[10] Bigram features were chosen for several reasons. First, in phonology, it is not practical to use substrings longer than bigrams or trigrams (Albright, 2009). While in principle, substring of any size can be used, due to sparsity issue, the number of logical combinations of ngrams would be large and a majority of the combinations will not be attested (Jurafsky & Martin, 2008). Since we would compute activation diversity over a trained AL lexicon which contains only 24 words, using trigrams would yield sparse and uninformative estimates of activation diversity. Second, trigram phonotactic information is already captured in a separate variable, consonant frame probability, as described later in this section, which directly captures the phonotactic pattern of identical consonants in C_C_ frames that spans across three adjacent segments.

nonwords *mamo* and *nuno* were computed for a miniature German lexicon consisting of three words *mami* (engl. mummy), *mama* (engl. mum) and *dodo* (engl. dodo). Outcomes are the three words and cues in our example are bigram segments.[11] The hypothetical test nonwords *mamo* and *nuno* are broken down to cues. To evaluate the activation level of each test nonword, we computed the sum of the weights in the vectors for each outcome.[12] Based on the present model, *mamo* receives a higher activation diversity than *nuno*.

Cues: \ Outcomes:	mami	mama	dodo
ma	1	1	0
am	1	1	0
mi	1	0	0
ma	0	1	0
mo	0	0	0
do	0	0	1
od	0	0	1
do	0	0	1
nu	0	0	0
un	0	0	0
no	0	0	0
sum of columns	2	2	0
Activation diversity *mamo*			4

Cues: \ Outcomes:	mami	mama	dodo
ma	1	1	0
am	1	1	0
mi	1	0	0
ma	0	1	0
mo	0	0	0
do	0	0	1
od	0	0	1
do	0	0	1
nu	0	0	0
un	0	0	0
no	0	0	0
sum of columns	0	0	0
Activation diversity *nuno*			0

Table 3: Illustration of how example test nonword *mamo* and *nuno* cues are checked against the matrix of the miniature lexicon.

Neighbourhood density While activation diversity has been shown to capture the same amount of variance as a number of psycholinguistic variables in psycholinguistic tasks, neighbourhood density was found to be the least correlated variable with activation diversity (Milin et al., 2017, **Table 5**). Neighbourhood density is a classic lexical variable that reflects analogical learning (Nosofsky, 1986). It is known to affect the speech production, the speech perception, and the acceptability judgement of nonwords. Nonwords with a high neighbourhood density are more accurate in nonword

[11] For illustration purposes, segments were used as cues. In the actual model, phonological features were used with the segmental bigrams being converted into featural bigrams. For instance, the featural bigram representation of the [+/–syllabic] feature for the first three segmental bigrams *ma*, *am*, and *mi* would be [-syllabic, +syllabic], [+syllabic, –syllabic], and [–syllabic +syllabic], which serve as separate cues. Note that [] is used to denote a sequence of featural properties across multiple segments rather than a single feature matrix. With N number of phonological features, a nonword of length L would yield $N \times (L\text{-}1)$ number of featural bigram cues.

[12] For illustration purposes, the activation values in the example were highly simplified to be either 1s or 0s.

repetition (Gathercole, 1995; Munson, Kurtz, & Windsor, 2005), take longer to reject in auditory lexical decision (Chuang et al., 2019; Luce & Pisoni, 1998), and are more acceptable as a real word in acceptability judgement (Bailey & Hahn, 2001; Harris, Neasom, & Tang, In prep) than nonwords with a low neighbourhood density. More specifically, in first and second language acquisition, words with a high neighbourhood density are better acquired than those with a low neighbourhood density (Coady & Aslin, 2003; Storkel, Armbrüster, & Hogan, 2006). Neighbourhood density is a measure of wordlikeness of a nonword. We therefore hypothesise that neighbourhood density of a nonword could influence whether it would get accepted as a word of the artificial language.

The GNM (Bailey & Hahn, 2001) was used to estimate the neighbourhood density of nonwords. GNM compares each nonword with all the words in the lexicon weighted by the phonological distances between each nonword and the real words and the token frequency of the real words.[13] The advantage of GNM over a strict one-segment difference neighbourhood density metric is that it takes into account the whole lexicon (not just words that are immediate neighbours) while adjusting for a neighbour's importance by how distant it is from the nonword (the more distant it is, the less important it is) and how frequent it is (the lower the frequency, the lower the importance). The gradient nature of GNM is particularly advantageous for computing over small lexicons (e.g., the AL lexicons) as otherwise many nonwords would have zero number of immediate neighbours. To determine the phonological distance between the nonwords and the real words, we first computed the phonetic similarities between all phones by taking the proportion of matched features over all features using the phonological distinctive feature system by PHOIBLE (Moran et al., 2014).[14] We then used these phonetic distances as weights to compute the weighted Levenshtein distance (Levenshtein, 1966) between two words.

For instance, to compute the neighbourhood density of a nonword, *mamo*, over a German lexicon, *mami*, which is a frequent word of German (engl.: mummy), will contribute more than *dodo* (engl.: dodo), which is an infrequent word of German. The reason to reward the nonword

[13] The GNM model proposed by Bailey and Hahn (2001) has a number of free parameters, which are obtained by post-hoc model fitting using regression. To reduce the number of parameters, we chose to use a monotonic frequency weighting scheme, as opposed to a quadratic frequency weighting scheme used in Bailey and Hahn (2001). Furthermore, we decided against using post-hoc model fitting to obtain the free parameters given the high computational demand; instead we opted to use existing published parameters by the study. The cost parameter for insertion and deletion was set to 0.7 and the sensitivity parameter was set to 1/0.1739.

[14] Our phonetic distance metric differs from the one used by the GNM model reported in Bailey and Hahn (2001). In the original model, distance was computed using a metric based on shared natural classes (Frisch, 1996), while our metric is a normalised number of feature mismatches. In a separate study by Bailey and Hahn (2005), they evaluated how well can a) feature mismatching, b) Frisch's metric, and c) empirical phoneme confusability in a number of word similarity judgement tasks, and they found that feature mismatching performed better than Frisch's metric. Given the findings of Bailey and Hahn (2005), we decided to employ a feature mismatching metric over Frisch's metric for our study. For Mandarin, the Chao tone digits was used to describe the tones (Tone 1: 55, Tone 2: 35, Tone 3: 21, Tone 4: 51) in terms of their tone levels, such that each tone is a vector of length two; the distance between two tones was the Euclidean distance of the two vectors.

mamo with a relatively high neighborhood density value is because the nonword *mamo* is phonologically closer to a frequent real word *mami* than to an infrequent real word *dodo*. This means the nonword *mamo* receives a strong neighborhood support by being phonologically close to a highly frequent real word *mamo*.

Consonant frame probability The experimental paradigm set out to test whether a consonant identity pattern is preferred over a non-identity pattern with lexical items of either the C_aVC_aV or the C_aVC_bV pattern. This could direct the participants' attention to the consonant frame C_C_, and therefore enhance the potential effect of lexical statistics over the consonant frames. For this reason, we included the probability of C_C_ as an additional lexical variable for examination. The probability of C_C_ is the number of word types with a given frame C_C_ divided by the number of all CVCV word types.[15] The choice of using type frequency over token frequency was to minimise the potential overlap with the activation diversity variable which was trained on token-based phonotactic information. Unlike the variables activation diversity and neighbourhood density, this was computed only with respect to the L1 lexicons (Brysbaert et al., 2011; Cai & Brysbaert, 2010) because the number of items with the same consonant frame was matched across frame types in the AL lexicons. Thus there is no variation with the probability of word types.

2.3.2. Language-independent variables

Three language-independent predictors were considered: Trial number, response time, and identity.

Trial Number Trial number is the number of the trial in a testing session. It was included to evaluate the effect of recency of the exposure of the artificial language, as we know that recently exposed grammatical structure are preferred (Luka & Barsalou, 2005).

We hypothesised that the participants would more likely accept a nonword as a word of the artificial language at the beginning of the testing session (a low trial number) than at the end of the session (a high trial number). In other words, as the experiment progressed (from a low trial number to a high trial number), the probability of a Yes response, which denotes accepting a nonword as a word of the artificial language, decreases.

Response Time Response time in this study is the time it takes to respond to each nonword. In many behavioural tasks response time is included as a control for the well-known trade-off between speed and accuracy (e.g., Davidson and Martin (2013); Heitz (2014)). This is not

[15] Our decision to use only CVCV words in the L1 lexicons is based on the fact that participants were exposed exclusively to CVCV nonwords as part of the ALL experiments. In principle, one could explore other estimates of consonant frame probability such as using all words, all sequential onsets, all CV.CV syllable sequences, or all CV.CV sequences at initials of words. We encourage the readers to examine other estimates in their individual studies.

how we should interpret response time because there is not a correct or an incorrect answer to whether a given nonword belongs to the artificial language exposed to the participants. We hypothesise that the longer the response time, the probability of a Yes response, which denotes accepting a nonword as a word of the artificial language, decreases.

Identity In the original study by Linzen and Gallagher (2017) in English, learners were shown to display an effect towards accepting more nonwords with an identical consonant pattern $C_{a-}C_{a-}$, than those with a non-identical consonant pattern $C_{a-}C_{b-}$ as words of the artificial language. Their learners were exposed to an artificial language with an overrepresentation of identical consonant frames relative to chance, and as a consequence learners exhibited a preference for identical consonant frames over non-identical consonant frames, which demonstrated learning of the phonotactic pattern. We expect learning of this identity pattern as well. Furthermore, a successful replication of this effect in our experiments would serve as indirect evidence that our experiments are successful replications of the original study in two new languages, thus lending validity to other findings from our study.

As in many ALL studies, Linzen and Gallagher (2017)'s learning effect, the identity effect, was tested for its generalisability. This was done by means by evaluating whether the identity effect holds in both familiar items (nonwords with consonant frames attested in the exposure phase) and unfamiliar items (nonwords with consonant frames not attested in the exposure phase). The evaluation of the identity effect in the unattested items is crucial since it enabled the researchers to investigate not only memory but also the generalisability of the observed patterns. Linzen and Gallagher (2017) found the identity effect was robust across both attested and unattested items and there was no significant difference between the two sets of items. The generalisability of the observed identity effect relies on the assumption that the unattested items do not resemble the exposure items in the consonant frame type; therefore, the observed identity effect is not due to the participant's experience of the exposure items. However, our AL variables and the time course variables in fact challenge this rigid assumption in two ways. First, while the unattested items did not overlap with the exposure items in their consonant frames at the segmental level, the same cannot be said at the individual featural level for both the consonant frames and indeed the vowels. Second, the time course variables evaluate the potential recency effect the participants have with respect to the exposure items. Therefore, we aim at jointly evaluating the identity effect with the AL and the time course variables. This would allow us to evaluate the role of exposure items in the processing of the unattested nonwords and see if the effect of identity would remain significant.

2.4. Model procedure

Linear mixed-effects logistic regression models were fit to the responses conducted using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015; R Core Team, 2013). For each of

the experiments, two models were fitted over the nonwords with attested C_C_ frames and the nonwords with unattested C_C_ frames respectively. The primary focus of our study is to evaluate whether or not lexical statistics such as the ones we outlined in Section 2.3 have an effect on the behavioural responses of an ALL experiment. To evaluate the effect of lexical statistics, four models were fitted over the German group's attested nonwords (Model 1) and unattested nonwords (Model 2), the Mandarin group's attested nonwords (Model 3) and the unattested nonwords (Model 4), predicting either a Yes or a No (base-level) response. In total, the four models were fitted with the predictor variables outlined in Section 2.3 as fixed effects and per-speaker and per-item random intercepts to allow for idiosyncrasies of individual speakers and items, as is typical of psycholinguistic research.

The regression structure of the initial models is shown below (note that the language-dependent predictors have either 'L1' or 'AL' in parentheses referring to their respective languages. L1 is referring to either German or Mandarin and AL is referring to an artificial language).

Response (Yes/No) ~ Identity + Activation diversity (L1) + Neighbourhood density (L1) + Activation diversity (AL) + Neighbourhood density (AL) + Consonant frame probability (L1) + Trial number + Response time + (1 | Participant) + (1 | Item)

Following standard practice in regression modelling, the continuous variables were z-score normalised (e.g., Baayen, 2008, Sec. 2.2). Z-score normalization allows us to compare the relative strength of our continuous predictors directly. The distributions of activation diversity and response time were skewed; therefore, they were log-transformed (base 10) before z-score normalization. Neighbourhood density was already on a log scale. Our categorical predictor Identity was sum-coded (Wissmann, Toutenburg, & Shalabh, 2007) with non-identical as the base level. The dependant variable, which is the participants' response (Yes/No), was coded with 'No' as the base level.

To evaluate potential collinearity issues, we computed the condition number (Belsley, Kuh, & Welsch, 1980), using the function *collin.fnc* in the library *languageR* (Baayen, 2013). Our variables have a condition number of 2.09 and 3.12 in Study I (German) and Study II (Mandarin) respectively. According to Baayen (2008), these values indicate no issues of collinearity.

These initial models were then simplified following a step-down, data-driven model selection procedure which compared nested models using the backward best-path algorithm of Gorman and Johnson (2013), making use of the *anova()* function and likelihood ratio test provided by *R*. After the step-down procedure, by-participant random slopes for the optimal fixed effects were fitted to ensure that our estimates of the effects of these factors will be relatively conservative. The random slopes were kept only if they are justified by the data and if there are no singularity or non-convergence issues. It has been suggested that the likelihood ratio test can be anti-conservative (Luke, 2017) when used as a measure of statistical significance. We, therefore,

chose a relatively liberal threshold of alpha, $\alpha = 0.1$, to be conservative in our model selection procedure, preferring to include potentially relevant predictors in the final model and their statistical significance will then be evaluated using bootstrapping. The final model will be presented in the next section.

The statistical significance of the individual predictors in all the models was evaluated by bootstrapping. Bootstrapping was carried out using the *bootmer* function in the *lme4* library. One-thousand bootstrap simulations were performed for each model. Bootstrapped p-values and confidence intervals at 95% were computed for each predictor in each model. We follow the conventional alpha-level of 0.05 for significance. Therefore, we will refer to any p-value below 0.05 as "significant" and any p-value greater than 0.05 but smaller than 0.1 as "near-significant".

In the following sections, we will first present the results of the *planned* analyses of Study I (German) and of Study II (Mandarin). In addition to these planned analyses, we will present the results of two *exploratory* analyses on the effect of attestedness and its interactions motivated by the results from Study I and Study II. Readers should note that these models are exploratory and their results should be interpreted with caution. This is because a) the effect of attestedness is not our study of interest, b) we have no *a-priori* reasons to examine the interactions of our many fixed effects and their random slopes, and c) we wanted to avoid making our models too complex to interpret and risk overfitting the small data that we have (approx. 1,000 data points per model).

3. Results

3.1. Study I: German

3.1.1. Study I: German: Attested

Table 4 summarises the fixed effects in Model 1, which is fitted over the German attested nonwords. The model structure of the best model is shown below.

Response (Yes/No) ~ Identity + Activation diversity (L1) + Neighbourhood density (L1) + Neighbourhood density (AL) + Consonant frame probability (L1) + Trial number + Response time + (1 + Response time | Participant) + (1 | Item)

We first examine the language-independent variables. All three of the language-independent variables were highly significant in the expected directions: These are Identity , trial number, and response time. The effect of identity was in the same direction as the original study for English, such that the nonwords with identical consonant were more likely to be accepted. The effect of trial number and response time suggested there is a recency effect – the higher the trial number and the longer the response time, the lower the likelihood of acceptance. In other words, the more recent a nonword was processed both across and within trials, the more likely it would get accepted. Having examined the language-independent variables unrelated to L1 and AL

effects, we move on to the three L1 variables and two AL variables. Two of the L1 variables were significant: Activation diversity and neighbourhood density. The third L1 variable, consonant frame probability, was near significant. All three L1 variables have a negative coefficient suggesting that the more L1-like the nonword (i.e., the higher the activation diversity, the higher the neighbourhood density, and the higher the consonant frame probability), the lower the likelihood of acceptance. Only one of the two AL variables was significant: Neighbourhood density. Activation diversity was not significant since it was dropped from the model selection. Neighbourhood density has a positive coefficient suggesting that the more AL-like the nonword (i.e., the higher the neighbourhood density), the higher the likelihood of acceptance. Finally, we compare the effect sizes of the L1 and AL variables. Neighbourhood density (AL) has a stronger effect than its L1 counterpart as well as the other two L1 variables, activation diversity and consonant frame probability. In sum, the influence of the AL lexicon is *stronger* than that of the L1 lexicon with the attested items. Random effects in Model 1 are summarised in Table 14 of the appendix.

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	1.4258	0.1223	11.6624	1.2098	1.6729	$<.001^{***}$
L1	Activation diversity	–0.1969	0.0893	–2.2044	–0.3684	–0.0288	$.016^{*}$
	Neighbourhood density	–0.2778	0.0935	–2.9708	–0.4641	–0.0985	$.004^{**}$
	Consonant frame prob.	–0.1483	0.0812	–1.8247	–0.3099	0.0017	$.072^{\cdot}$
AL	Activation diversity	–	–	–	–	–	*n.s.*
	Neighbourhood density	0.4568	0.1258	3.6303	0.2162	0.7012	$<.001^{***}$
Language independent	Identity (Identical vs. Non-identical)	0.6749	0.1553	4.3445	0.3847	0.9761	$<.001^{***}$
	Trial number	–0.5205	0.0717	–7.2538	–0.6644	–0.3882	$<.001^{***}$
	Response time	–0.6544	0.0940	–6.9574	–0.8385	–0.4771	$<.001^{***}$
Number of observations: 1856; number of participants: 232; number of items: 128							
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

Table 4: Fixed effects summary for Study I (German, attested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

3.1.2. Study I: German: Unattested

Table 5 summarises the fixed effects in Model 2, which is fitted over the German unattested nonwords. The model structure of the best model turned out to be the same as in Model 1 (see Section 3.1.1).

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	1.3606	0.1156	11.7709	1.1423	1.6034	$<.001^{***}$
L1	Activation diversity	–0.3528	0.0971	–3.6329	–0.5431	–0.1692	$<.001^{***}$
	Neighbourhood density	–0.3232	0.0964	–3.3517	–0.5123	–0.1362	$<.001^{***}$
	Consonant frame prob.	–0.2944	0.0884	–3.3306	–0.4729	–0.1246	$<.001^{***}$
AL	Activation diversity	–	–	–	–	–	$^{n.s.}$
	Neighbourhood density	0.2246	0.0938	2.3939	0.0363	0.4128	$.012^{*}$
Language independent	Identity (Identical vs. Non-identical)	0.5464	0.1635	3.3416	0.2338	0.8734	$<.001^{***}$
	Trial number	–0.4082	0.0660	–6.1819	–0.5425	–0.2804	$<.001^{***}$
	Response time	–0.4951	0.0808	–6.1263	–0.6581	–0.3474	$<.001^{***}$
Number of observations: 1856; number of participants: 232; number of items: 128							
Level of significance: · (p ⩽ 0.1), * (p ⩽ 0.05), ** (p ⩽ 0.01), *** (p ⩽ 0.001).							

Table 5: Fixed effects summary for Study I (German, unattested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

We first examine the language-independent variables. Like Model 1, Identity, trial number and response time were highly significant in the expected directions. Identical consonant frames are accepted more than non-identical consonant frames; the higher the trial number, the lower the likelihood of acceptance; and the longer the response time, the lower the likelihood of acceptance. Therefore, the effect of identity and recency were found in both attested and unattested nonwords. We now examine the three L1 variables and two AL variables. All three of the L1 variables were highly significant in the same direction as Model 1: Activation diversity, neighbourhood density, and consonant frame probability. Like Model 1, the more L1-like the

nonword (i.e., the higher the activation diversity, the higher the neighbourhood density, and the higher the consonant frame probability), the lower the likelihood of acceptance. Just like Model 1, only neighbourhood density (AL) was significant with a positive coefficient, while activation diversity (AL) was dropped from the model selection. Again we compare the effect sizes of the L1 and AL variables. Contrary to Model 1, neighbourhood density (AL) has a *weaker* effect than its L1 counterpart as well as the other two L1 variables, activation diversity and consonant frame probability. In sum, the influence of the AL lexicon is *weaker* than that of the L1 lexicon with the unattested items. Random effects in Model 2 are summarised in Table 15 of the appendix.

Figure 2 summarises the effect size of the significant variables in Model 1 and Model 2. It illustrates that the direction of the significant effects is consistent across the attestedness condition. L1 variables have a negative effect while the AL variable has a positive effect. Identical consonant frames are more likely to be accepted. Nonwords that were responded to earlier in the experiment are more likely to be accepted across and within trials.

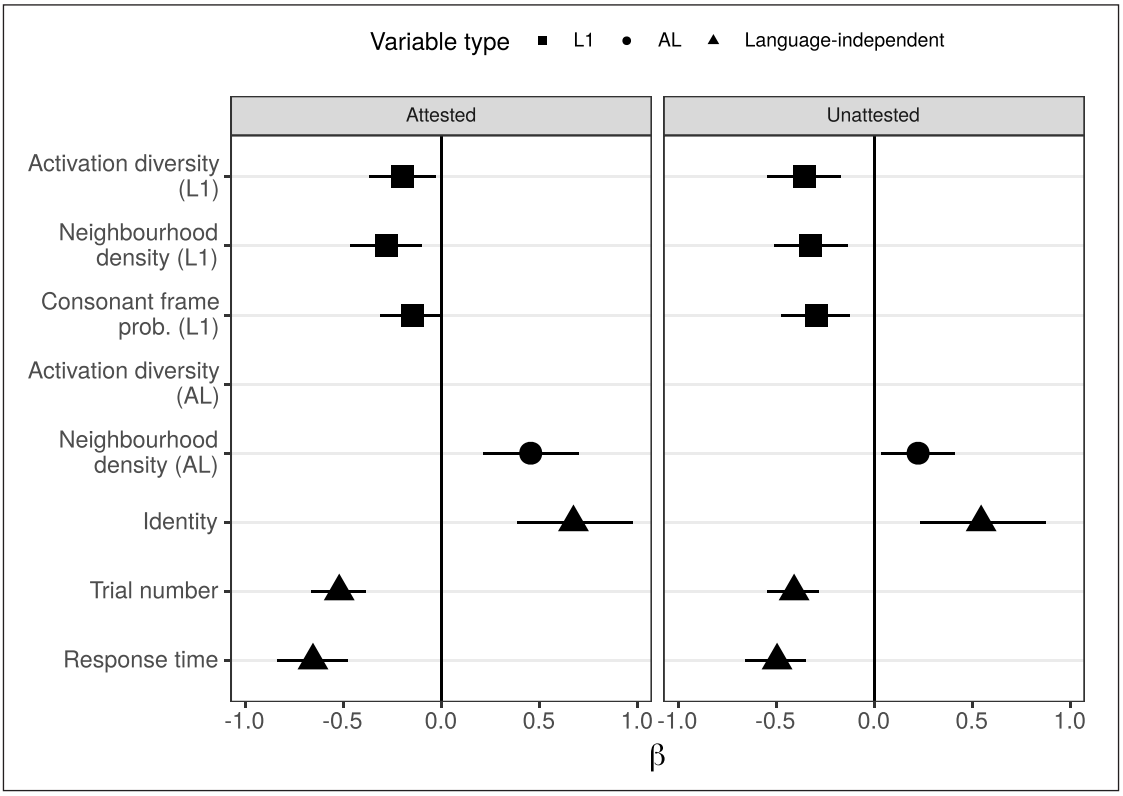

Figure 2: German: Effect size with a 95% confidence interval by attestedness (Model 1: Attested, Model 2: Unattested) and variable type (■: L1, ●: AL, ▲: Language-independent); variables not illustrated were dropped during model selection.

3.2. Study II: Mandarin

3.2.1. Study II: Mandarin: Attested

Table 6 summarises the fixed effects in Model 3, which is fitted over the Mandarin attested nonwords. The model structure of the best model is shown below.

Response (Yes/No) ~ Identity + Neighbourhood density (L1) + Neighbourhood density (AL) + Trial number + Response time + (1 + Response time | Participant) + (1 | Item)

Again we first examine the language-independent variables. Like the German group (Model 1 and Model 2), Identity, trial number, and response time were highly significant in the expected directions. Identical consonant frames are accepted more than non-identical consonant frames; the higher the trial number, the lower the likelihood of acceptance; and the longer the response time, the lower the likelihood of acceptance.

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	0.9209	0.0948	9.7063	0.7462	1.1172	$<.001^{***}$
L1	Activation diversity	–	–	–	–	–	*n.s.*
	Neighbourhood density	–0.2405	0.0793	–3.0309	–0.4025	–0.0867	$.002^{**}$
	Consonant frame prob.	–	–	–	–	–	*n.s.*
AL	Activation diversity	–	–	–	–	–	*n.s.*
	Neighbourhood density	0.3226	0.1006	3.2079	0.1118	0.5267	$.002^{**}$
Language independent	Identity (Identical vs. Non-identical)	0.7173	0.1349	5.3176	0.4647	0.9859	$<.001^{***}$
	Trial number	–0.2774	0.0609	–4.5531	–0.3980	–0.1581	$<.001^{***}$
	Response time	–0.3651	0.0759	–4.8140	–0.5146	–0.2266	$<.001^{***}$
Number of observations: 1752; number of participants: 219; number of items: 128							
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

Table 6: Fixed effects summary for Study II (Mandarin, attested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

Unlike the German group (Model 1), of the three L1 variables, only neighbourhood density was significant, while both activation diversity and consonant frame probability had no significant effect on nonword acceptance. The effect of neighbourhood density (L1) suggests that the more L1-like the nonword, the lower the likelihood of acceptance. Of the AL variables, only neighbourhood density AL was significant in the same direction – the more AL-like the nonword, the higher the likelihood of acceptance. Activation diversity (AL) was not significant. Finally, the effect size of neighbourhood density (AL) is higher than its L1 counterpart. Besides the insignificance of the two L1 variables, activation diversity and consonant frame probability, the direction of the L1 and AL effects and their relative difference of their effect sizes are consistent with that of the German group (Model 1). In sum, the influence of the AL lexicon is *stronger* than that of the L1 lexicon with the attested items. Random effects in Model 3 are summarised in Table 16 of the appendix.

3.2.2. Study II: Mandarin: Unattested

Table 7 summarises the fixed effects in Model 4, which is fitted over the Mandarin unattested nonwords. The model structure of the best model is shown below.

> Response (Yes/No) ~ Identity + Neighbourhood density (L1) + Consonant frame probability (L1) + Neighbourhood density (AL) + Trial number + Response time + (1 | Participant) + (1 | Item)

The language-independent variables were all significant and their effects are in the same direction as all the other models. Identity, trial number and response time were highly significant in the expected directions.

Of the three L1 variables, activation diversity was not significant and was dropped during model selection, while neighbourhood density and consonant frame probability were both significant with a negative effect. The more L1-wordlike the nonword, the lower its likelihood of acceptance. Of the two AL variables, only neighbourhood density was significant with a positive effect. A comparison of the effect sizes across the L1 and AL variables suggest that neighbourhood density (AL) has a weaker effect than its L1 counterpart as well as consonant frame probability (L1).

Besides the insignificance of the L1 variable activation diversity, the direction of the L1 and AL effects and the relative difference of their effect sizes are consistent with that of the German group (Model 2). Again, the influence of the AL lexicon is *weaker* than that of the L1 lexicon with the unattested items. Random effects in Model 4 are summarised in Table 17 of the appendix.

Figure 3 summarises the effect size of the significant variables in Model 3 and Model 4. It illustrates that the direction of the significant effects is consistent across the attestedness condition with a minor difference in how consonant frame probability (L1) is only significant in the unattested condition.

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	1.1052	0.1118	9.8829	0.8917	1.3261	$<.001^{***}$
L1	Activation diversity	–	–	–	–	–	*n.s.*
	Neighbourhood density	–0.2596	0.0964	–2.6928	–0.4424	–0.0806	$.008^{**}$
	Consonant frame prob.	–0.2507	0.0952	–2.6330	–0.4298	–0.0662	$.004^{**}$
AL	Activation diversity	–	–	–	–	–	*n.s.*
	Neighbourhood density	0.2147	0.0938	2.2861	0.0307	0.3975	$.016^{*}$
Language independent	Identity (Identical vs. Non-identical)	0.8441	0.1701	4.9618	0.5122	1.1841	$<.001^{***}$
	Trial number	–0.2754	0.0639	–4.3086	–0.4028	–0.1508	$<.001^{***}$
	Response time	–0.4418	0.0743	–5.9449	–0.5851	–0.2889	$<.001^{***}$
Number of observations: 1752; number of participants: 219; number of items: 128							
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

Table 7: Fixed effects summary for Study II (Mandarin, unattested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

3.3. Attestedness and its interactions

In the models for attested items (model 1, and model 3), we observed that AL variables have a stronger effect than the L1 variables. In contrast, in the models for the unattested items (model 2, and model 4), the L1 variables have a stronger effect than the AL variables. To better evaluate this pattern, we conducted two additional analyses by testing if attestedness interacts with L1 and AL variables for each language group separately. Attestedness was sum-coded with unattestedness as the base level. Below, we present the best model for German; next we present the best model for Mandarin.

> Response (Yes/No) ~ Identity + Activation diversity (L1) + Neighbourhood density (L1) + Neighbourhood density (AL) + Consonant frame probability (L1) + Trial number + Response time + Activation diversity (L1):Attestedness + Neighbourhood density (AL):Attestedness + Consonant frame probability (L1):Attestedness + (1 + Response time | Participant) + (1 | Item).

Response (Yes/No) ~ Identity + Neighbourhood density (L1) + Consonant frame probability (L1) + Neighbourhood density (AL) + Trial number + Response time + Consonant frame probability (L1):Attestedness + (1 + Response time | Participant) + (1 | Item)

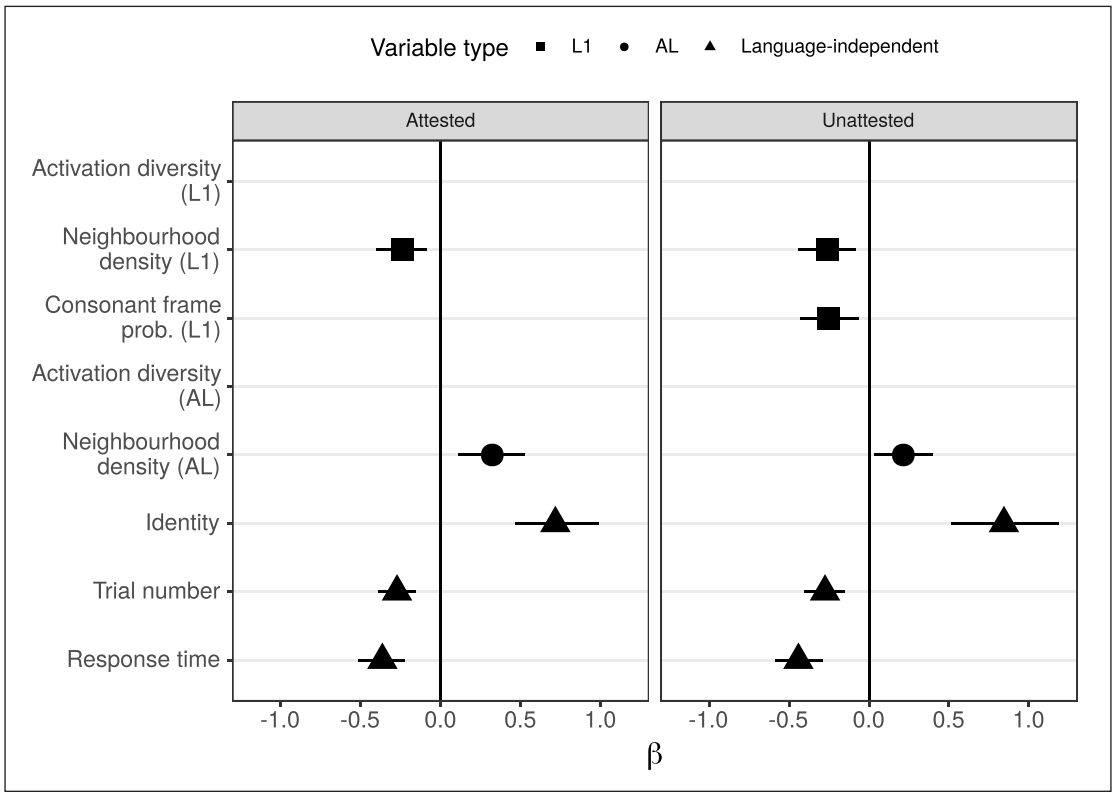

Figure 3: Mandarin: Effect size with a 95% confidence interval by attestedness (model 3: Attested, model 4: Unattested) and variable type (■: L1, ●: AL, ▲: Language-independent); variables not illustrated were dropped during model selection.

Table 8 and **Table 9** summarise the fixed effects of the two regression models, which are fitted over the German nonwords and the Mandarin nonwords respectively. There is no main effect of attestedness found in either the German group or the Mandarin group. For German, attestedness interacts with neighbourhood density (AL) significantly and two interaction terms were near significant: Activation diversity (L1), and consonant frame probability (L1). Attestedness did not significantly interact with neighbourhood density (L1), and it was dropped during model selection. For Mandarin, only one of the four interaction terms was significant: Attestedness interacts with consonant frame probability (L1). Attestedness did not significantly interact with neighbourhood density (L1) and neighbourhood density (AL), and they were dropped during model selection. Random effects of the two models are summarised in Table 18 and Table 19 of the appendix.

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	1.4229	0.1033	13.7782	1.2352	1.6254	$<.001^{***}$
L1	Activation diversity	–0.2822	0.0730	–3.8654	–0.4301	–0.1388	$.002^{**}$
	Neighbourhood density	–0.2911	0.0738	–3.9476	–0.4328	–0.1456	$<.001^{***}$
	Consonant frame prob.	–0.2418	0.0676	–3.5751	–0.3776	–0.1095	$<.001^{***}$
AL	Activation diversity	–	–	–	–	–	$^{n.s.}$
	Neighbourhood density	0.3151	0.0796	3.9600	0.1547	0.4698	$<.001^{***}$
Language independent	Identity (Identical vs. Non-identical)	0.6207	0.1266	4.9020	0.3789	0.8749	$<.001^{***}$
	Trial number	–0.4811	0.0480	–10.0181	–0.5750	–0.3860	$<.001^{***}$
	Response time	–0.6254	0.0940	–8.7904	–0.7635	–0.4908	$<.001^{***}$
Attestedness and its interactions	Attestedness (AL) (Attested vs. Unattested)	0.0238	0.1037	0.2295	0.1700	0.2187	$.792^{n.s.}$
	Attestedness (AL): Activation diversity (L1)	0.2097	0.1061	1.9764	–0.0072	0.4284	$.058^{\cdot}$
	Attestedness (AL): Neighbourhood density (L1)	–	–	–	–	–	$^{n.s.}$
	Attestedness (AL): Consonant frame prob. (L1)	0.1856	0.0949	1.9550	-0.0099	0.3777	$.054^{\cdot}$
	Attestedness (AL): Neighbourhood density (AL)	0.2994	0.1284	2.3318	0.0340	0.5438	$.014^{*}$
Number of observations: 3712; number of participants: 232; number of items: 128							
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

Table 8: Fixed effects summary for Study I with attestedness and its interactions with L1 and AL variables (German, attested, and unattested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

		β	SE	z	$CI_{Lower95\%}$	$CI_{Upper95\%}$	$p_{Bootstrapped}$
	(Intercept)	1.0660	0.0897	11.8720	0.8879	1.2534	$<.001^{***}$
L1	Activation diversity	–	–	–	–	–	$^{n.s.}$
	Neighbourhood density	–0.2400	0.0687	–3.4948	–0.3727	–0.1075	$<.001^{***}$
	Consonant frame prob.	–0.1592	0.0744	–2.1409	–0.3126	–0.0111	$.036^{*}$
AL	Activation diversity	–	–	–	–	–	$^{n.s.}$
	Neighbourhood density	0.2524	0.0698	3.6153	0.1101	0.3972	$<.001^{***}$
Language independent	Identity (Identical vs. Non-identical)	0.8130	0.1306	6.2258	0.5610	1.0732	$<.001^{***}$
	Trial number	–0.2769	0.0438	–6.3300	–0.3621	–0.1891	$<.001^{***}$
	Response time	–0.4532	0.0631	–7.1792	–0.5685	–0.3367	$<.001^{***}$
Attestedness and its interactions	Attestedness (AL) (Attested vs. Unattested)	–0.1312	0.0928	–1.4135	–0.3351	0.0590	$.208^{n.s.}$
	Attestedness (AL): Activation diversity (L1)	–	–	–	–	–	$^{n.s.}$
	Attestedness (AL): Neighbourhood density (L1)	–	–	–	–	–	$^{n.s.}$
	Attestedness (AL): Consonant frame prob. (L1)	0.2051	0.0858	2.3894	0.0303	0.3820	$.030^{*}$
	Attestedness (AL): Neighbourhood density (AL)	–	–	–	–	–	$^{n.s.}$
Number of observations: 3504; number of participants: 219; number of items: 128							
Level of significance: · ($p \leqslant 0.1$), * ($p \leqslant 0.05$), ** ($p \leqslant 0.01$), *** ($p \leqslant 0.001$).							

Table 9: Fixed effects summary for Study II with attestedness and its interactions with L1 and AL variables (Mandarin, attested, and unattested). β: Coefficient; SE: Standard error; z: z-value; $CI_{Lower95\%}$ and $CI_{Upper95\%}$: 95% confidence intervals of the coefficient from bootstrapping; $p_{Bootstrapped}$: p-value from bootstrapping simulations.

4. Discussion

In order to find out whether a learning meachanism which has been demonstrated in English speakers is universal, the present study aimed at investigating two distinct populations of speakers, namely Mandarin and German. In the present paper we utilize a new approach on how to disentangle (possibly universal) effects of learning from effects of the native language of the speakers and from effects of the artificial language designed for the experiment. We did so by statistically controlling for wordlikeness by means of analogical and discriminative modeling. Model outputs enabled us to control for L1 as well as for AL effects. In the discussion we go through the observed effects step by step. We start with the original learning effect. We then briefly discuss some language-independent task effects. After that we turn to L1 effects and then AL effects and the comparison of the two. We close with a discussion about our findings and implications for future research on learning behaviour.

4.1. The consonant identity effect

Linzen and Gallagher (2017) have shown that English speakers were able to rapidly learn a consonant identity pattern from an artificial language and the present study replicates these findings for German and Mandarin: In all models identity showed up as a highly significant factor – nonwords with identical consonsants are more likely to be accepted as part of the new language than nonwords with non-identical consonants. Our study confirms the robust and strong identity effect for German and Mandarin speakers (**Figure 4**), two populations which are very distinct from each other. The figure demonstrates that the identity effect is robust as it shows up in both population regardless of whether the nonword was attested or unattested during training.

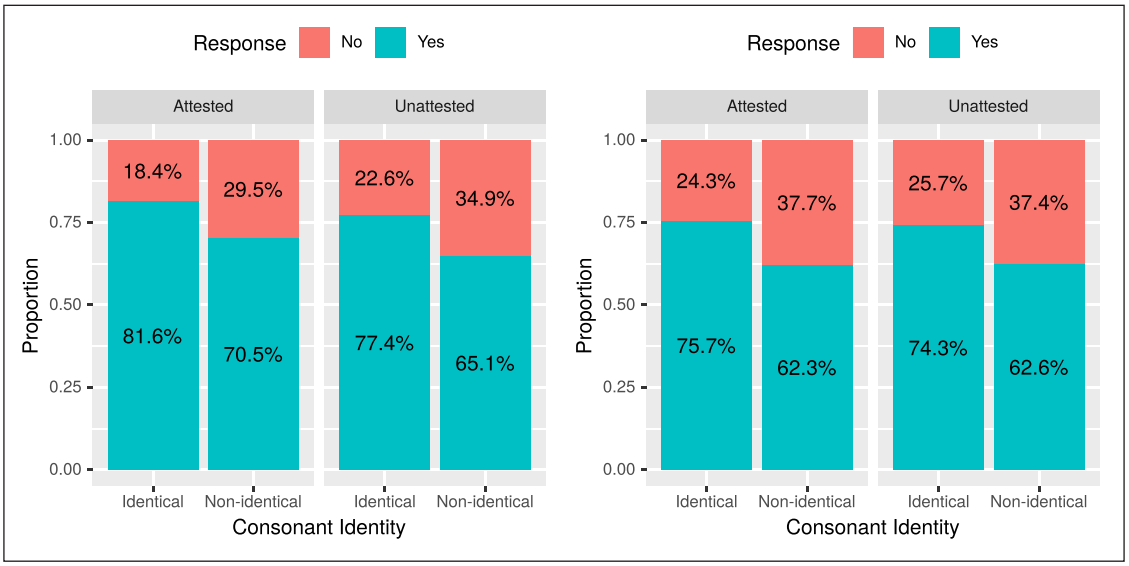

Figure 4: Response distribution by identity and attestedness in the German group (left) and the Mandarin group (right).

This is remarkable for two reasons: First, it seems likely that a general cognitive mechanism promoting learning of identity patterns also affects smaller units such as features and is responsible for the relative learnability of harmony patterns (Baer-Henney & van de Vijver, 2012; Finley, 2012; Martin & White, 2021; Moreton, 2012; Wilson, 2003). The identity effect seems not to be a language-specific phenomenon but it seems to be universal as has been proven by demonstrating it in three distinct languages.

Second, replication of findings in science has gained increasing attention over the last few years. Some scientific subfields including linguistics discuss a so-called replication crisis since replicability rates for empirical studies have been found to be very low (Nieuwland et al., 2018; Stack, James, & Watson, 2018). This problems calls for convincing accumulating evidence for the effects we report (Nicenboim, Roettger, & Vasishth, 2018; Roettger & Baer-Henney, 2019), and the present paper allows us to show valuable additional evidence for an effect that has been reported for English speakers and has replicated the original effect in German and Mandarin speakers.

It is worth noting that the identity effect in our study is strong enough not to be masked by our additional L1 and AL variables. In other words, the effect cannot be attributed to language-specific characteristics of either the artificial language used in the experiment or the L1 of the speakers. The present case hence illustrates a scenario in which an effect under investigation is robust enough not to be influenced by language-specific variables.

4.2. Task effects

While task effects do not play a central role in the present paper we still acknowledge the fact that to the best of our knowledge, the examination of task effects has been overlooked by most ALL researchers. In the present study we controlled for trial number and response time. We observed that the acceptance rate decreased within and across trials irrespective of language and attestedness. Longer response times resulted in more negative responses and so did higher trial numbers. Both types of artificial CVCV items (identical and non-identical frames) were more likely to be rejected as time proceeded within and across trials. We cannot confidently provide a singular interpretation of these task effects, since there are many possible explanations. Later trials in the experiment may suffer from fatigue or boredom of the participants or the additive exposure to stimuli during the test phase as compared to stimuli from the training had an effect (e.g., resulting in a tendency to accept items that are more similar to accepted test stimuli). That longer reaction times come with more negative responses could be a problem of inattention or decision difficulty.

The task effects are an important side finding, since it clearly suggests that even for a rapid learning paradigm of ALL (in our case, with only 16 test trials), task effects already play a role in shaping the rate of learning. Furthermore, the effect size of the task effects was consistently higher than the L1 and the AL variables across all models, which further highlights their importance.

Future studies may take a closer look at task effects, for instance, by including additionally collecting measures of memory, fatigue, boredom, and attention.

4.3. L1 effects

We have found that L1 variables, that is activation diversity, neighbourhood density and consonant probabilities of types, all have a negative effect in ALL experiments. For Germans we found that (1) the higher its NDL activation diversity is, (2) the higher its neighbourhood density is, and (3) the higher its consonant frame probability is, the less likely a test item is to be accepted as a word of the artificial language. For German, all three L1 variables were significant with both the attested and the unattested nonwords. For Mandarin speakers we found that (1) the higher its neighbourhood density is, and (2) the higher its consonant frame probability is (only unattested), the less likely a test item is to be accepted as a word of the artificial language. For Mandarin, activation diversity was not significant with both the attested and the unattested nonwords, and consonant frame probability was only significant with the unattested nonwords. Taken together, the more similar an item is to L1, the more likely it is to be rejected. Let us first discuss why there is a negative effect of L1 knowledge.

Learners in our experiment faced a short exposure phase in which they were told to listen to words of a new language. Their test task later on was to judge whether newly encountered words could also be part of that language. We speculate that participants may have suppressed their L1 knowledge in order to learn the AL, and this suppression was reinforced by the design of the experiment. The participants were told that AL was a *new* language that they would learn; therefore, it was safe for the participants to assume that this new language was distinct from their L1. Everything that reminded participants of their L1 could have been suppressed and, as a consequence, rejected: The more similar an item was to L1, the less likely it was to be accepted as a word of AL.

Next, a few speculations can be made as to why the L1 effects were weaker in Mandarin than in German. Note that the direct comparisons of the two groups in a single statistical model were not conducted. Therefore, these speculations should be taken with caution. First, the Mandarin concept of a word is different. The vast majority of morphemes in Mandarin Chinese consist of single syllables (Norman, 1988). Mandarin is also a compounding language. Most disyllabic words, therefore, consist of syllables, which are also morphemes, and these morphemes are highly salient as they can serve as individual words (Zhou & Marslen-Wilson, 1994). The same is not true for German disyllabic words since most German syllables are not morphemes that can stand alone as monomorphemic words. Therefore, compared to the German participants, Mandarin participants could readily assign meanings to their disyllabic nonwords. In fact, recent studies of the lexical processing of *English* nonwords have already shown that nonwords do interact with the semantic space (Chuang et al., 2019; 2020) and so does the production of English nonwords

(Schmitz, Plag, Baer-Henney, & Stein, 2021). We speculate that the role semantics plays in the nonword processing will be a lot larger for Mandarin than for English or German. The weaker L1 effects of Mandarin could be due to how our L1 measures do not take semantics into account, which is likely to play a strong role for Mandarin.

Second, Mandarin is shown to have productive reduplication patterns, affecting major lexical categories (verbs, adjectives, and nouns) (F.-Y. Chen, Mo, Huang, & Chen, 1992; Melloni & Basciano, 2018). For instance, reduplication can take place with both monosyllabic bases (A as AA), and disyllabic bases (AB as ABAB). Our lexicon is unlikely to contain as many reduplicated forms as speakers might actually use. As our lexicon was estimated using a corpus of texts, it suffers from a data sparsity issue. Low frequency reduplicated forms will not be attested unless a larger corpus is used (Blevins, Milin, & Ramscar, 2017).[16] This could lead to a poor estimate of our L1 variables as they were computed over a lexicon with an under-representation of reduplicated forms. The adverse effects of this under-representation are likely to be amplified by how the nonwords with identical consonant frames resemble reduplicated forms. The weaker L1 effects could simply be an artefact of our lexicon.

4.4. AL effects

We have found that one of the two AL variables, neighbourhood density, has a positive effect in both ALL studies. The German and Mandarin speakers were more likely to accept a test nonword as a word of the artificial language if it had a higher AL neighbourhood density. The other AL variable, activation diversity, had no effect in either language group. For both language groups, neighbourhood density was significant with both the attested and the unattested nonwords. We found that AL effects are weaker than the L1 effects in unattested items. In sum, the more similar an item is to AL, the more likely it is to be accepted. In other words, participants were doing their task as expected, such that they were judging incoming new test nonwords based on their AL wordlikeness.

Why were participants sensitive only to the AL neighbourhood density measure but not to the AL activation diversity measure in all of the models? We speculate that this asymmetry in the two AL variables lies in how well the exposure items were encoded and the robustness of the two lexical modelling approaches.

[16] This issue of data sparsity can be seen with the lexical statistical analyses conducted in Boll-Avetisyan (2012, Ch. 4). The study used a Mandarin Chinese lexicon containing phonological transcriptions of telephone conversations (266,642 word tokens) to estimate the degree of consonant co-occurrences with shared [place] in CVC sequences. They repeated their analyses with and without any attested reduplicated forms and they concluded that there is no strong evidence for constraints regarding PLACE on non-adjacent consonants. Only 0.33% of the lexicon were identified as reduplicated forms (901 out of 266,642 word tokens), which is likely an underestimation of the productive reduplication patterns in Mandarin. This underestimation could influence the reliability of their lexical analyses.

We have reasons to believe that the participants did not remember the exposure items correctly. In other words, the exposure items were not well encoded in memory as distinct lexical items and not encoded faithfully to the input. First, the instructions did not ask the participants to memorise these exposure items, and they were told that the subsequent task would involve only *new* words; therefore, they had no reason to learn these exposure items. Second, there were no tasks that required them to recall these exposure items. Third, in natural language, word learning is driven by the need to distinguish different meanings for communication. However, unlike real words, these exposure items are void of meaning[17]. Let us now examine how poor item encoding would have an effect on how well the two models estimate wordlikeness.

The naive discriminative learning (NDL) model aims to discriminate between a set of outcomes given their corresponding cues. The effect of inaccurate encoding of the exposure items could severely affect how well the naive discriminative learner predicts the performance of our human learners because the activation diversity measure is a function of the amount of activation of cues across all outcomes. If the exposure items were incorrectly encoded[18], hence stored as incorrect outcomes, the cues that correspond to the incorrect outcomes would not be represented as originally intended and the cues that correspond to the intended outcome would be erroneously strengthened. As a consequence, the cues of the test nonwords would not find their actual activation weights in the intended cue-to-outcome matrix. In this way, an inaccurate encoding would cause the NDL model to represent the phonological cues that correspond to the outcomes differently from our human participants.

What about the AL neighbourhood density measure? We speculate that the GNM is robust to the precise encoding of the phonological forms, which constitute the AL lexicon. GNM measures the overall wordlikeness based on the frequency-weighted sum of the phonological distances between the target word and every word in the lexicon, and crucially these phonological distances are independent of each other. While the incorrect encoding of an exposure item in the AL lexicon would still affect the phonological distance between itself and the test nonword, it would not affect how other items contribute to the overall wordlikeness score of the target word. This is what makes GNM different from NDL. In NDL, the incorrect encoding of an exposure item would affect the activation values of the intended cues and incorrect cues which are shared by other outcomes. Thus, even if some of the exposure words were incorrectly encoded, the effect on GNM would be smaller than that on NDL.

Finally, there is the possibility that NDL requires more trials to get a more reliable cue-to-outcome matrix, and that it is not suitable for modelling rapid learning in contrast to the

[17] While speakers might assign meanings to the nonwords based on their similarity to real words (e.g., Bailey & Hahn, 2005; Schmitz et al., 2021), the nonwords themselves are *inherently* void of meaning.

[18] An item might be incorrectly encoded because of speech misperception (e.g., Bennett, Tang, & Ajsivinac Sian, 2018; Miller & Nicely, 1955; Nieder & Tang, 2023; Tang, 2015; Tang & Nevins, 2014) or misremembering.

probabilistic model by Linzen and O'Donnell (2015). Linzen and O'Donnell (2015)'s model has the ability to represent different degrees of generalisation which makes it suitable for rapid learning. It was demonstrated that, together with a parsimony bias, their model was able to simulate learning patterns similar to human learners in Linzen and Gallagher (2017), specifically in how human learners favour rapid abstraction. Furthermore, NDL makes the assumption that the association strengths between cues and outcomes will no longer change (i.e. the model has reached the adult state of learning) (Danks, 2003). Given the small number of trials in the exposure phase, this might be too big an assumption to make.

4.5. A comparison of L1 and AL effects

We found an asymmetric pattern in the strength of L1 and AL predominantly in the German group: While AL variables are stronger than L1 variables in attested items, L1 variables are stronger than AL variables in unattested items (see Section 3.3). This can be explained by how wordlike the nonwords were with respect to the L1 and the AL lexicons. In comparison with the unattested items, the attested items were even more similar to the AL lexicon than to the L1 lexicon since the consonant frames of the attested items are the same as all of the items in the AL lexicon, with only the vowels being different. Likewise, L1 has more impact than AL on the unattested items, because the participants are more uncertain about these items being a word of AL. This uncertainty causes the participants to rely on what their existing linguistic knowledge, namely their L1. Therefore, the participants associated the attested nonwords more strongly with the AL than the L1, thus the AL effects were enhanced and the L1 effects were weakened.

There is no main effect of attestedness found in either the German group or the Mandarin group. Hence, the population difference cannot be explained on the basis of the Mandarin speakers not being able to distinguish attested items from the unattested items. Instead, we believe that this is more likely due to the L1 effects being weaker in the Mandarin group in general, which we speculated to be caused by the peculiarities of Mandarin as discussed in Section 4.3.

4.6. Our findings and implications for research on learning behaviour

Our experiment showed that the original learning effect could be shown to be present in German and Mandarin speakers and, therefore, we could confidently argue for its language-independence and universality. We have moreover shown that there are language-independent task effects but crucially that also language-specific effects are at work: More L1 wordlikeness leads to more acceptance of a stimulus while more AL wordlikeness leads to more rejections of a stimulus. These findings result in implications for future research on learning behaviour: Our findings show that language learning research needs stricter design control of the miniature training language than commonly assumed. With the present paper we aim to show how ALL research but also other learning research can be refined methodologically.

Our test case shows how the lexicons influence decisions in an ALL experiment with acceptability judgements, specifically an experiment which examines learning of a probabilistic phonotactic identity pattern (Linzen & Gallagher, 2017). There is reason to believe that lexical effects will also influence ALL studies in which other methods are evaluated, including the whole range of perceptual and productive test procedures.

Our findings can contribute beyond the investigation of phonotactic learning. One strand of research of special interest to the proposed approach is that of language learning biases. Biases are not only seen as driving forces in language learning; they are said to shape human language. They are seen as universal strategies and thus call for cross-linguistic studies that can justify broad implications based on these biases. Such cross-linguistic studies are methodologically challenging and we offer a solution for this problem. Biases need more such experimental investigations as some of them are controversially discussed. There is a relative consensus about the *complexity* bias, which says that featurally simple phonological patterns are learned more easily than more complex patterns (Moreton & Pater, 2012b), while the *substantive* bias, a bias that favours phonetically motivated patterns (Baer-Henney et al., 2015; DeMille et al., 2018; Finley, 2012; Glewwe, 2019; Greenwood, 2016; Martin & White, 2021; Moreton & Pater, 2012a; Tang, DeMille, Frijters, & Gruen, 2020; Wilson, 2006) is heavily debated. Other examples are the *locality* bias showing a preference for local over non-local patterns (Baer-Henney & van de Vijver, 2012; McMullin & Hansson, 2016; Tang & Akkuş, 2022; White et al., 2018), or the *regularisation* bias showing a preference for regular over irregular patterns (Fehér, Wonnacott, & Smith, 2016; Hudson Kam & Newport, 2005; Nevins, Rodrigues, & Tang, 2015; Samara, Smith, Brown, & Wonnacott, 2017; Smith et al., 2017; Smith & Wonnacott, 2010; Tang & Nevins, 2013). One essential problem to date is that most evidence for universal biases comes from investigations of only a few languages. Most of the abovementioned studies report evidence from one language each despite the fact that the claim for universality calls for crosslinguistic studies. The present data highlights the risk of underestimating the role of lexical statistics in ALL research. As we have argued, the influence of a bias could be masked by lexical effects. With the present approach we now offer a solution to these problems and show that lexical effects can be modelled, which comes with several additional advantages.

First, the present approach facilitates carrying out cross-linguistic studies. As with the current study, one could analyse each language group separately and evaluate if the learning bias persists even when language-specific factors (as captured by our L1 and AL variables) are taken into account. Furthermore, one could evaluate the relative strength of the learning bias to see if it is similar across the two language groups by regressing both language groups together. This is particularly advantageous when we would like to compare languages that have relatively distinct inventories. By controlling for lexical variables we no longer need one set of items for all the languages under investigation and we are no longer constrained by phoneme inventory overlap as in White et al. (2018). It allows us to investigate effects of the language groups under

investigation in much more detail than in common practice. Crucially, lexical variables can be used to control two types of L1 similarities of an AL test item. An AL test item can be similar to the L1 and can be supported by L1 lexical statistics in terms of a) the target AL pattern of interest, and b) any other non-target patterns. Typically, ALL experimenters would avoid testing a population whose L1 (partially) contains the target AL pattern of interest. The reason is that the amount of learning of a target AL pattern is expected to be greater with languages with a higher similarity of the target AL pattern. However, by directly modelling the degree of L1 similarity with respect to the target AL pattern of interest, researchers would be able to to test a target AL pattern even in languages that contain the pattern of interest. The present account offers this unique advantage of being able to investigate the same learning mechanisms in multiple and distinct languages.

Second, our approach would be beneficial for second language acquisition (SLA) research. While it is usually discussed independently from ALL learning, it is methodologically and thematically related. SLA can be seen as a real life setting of ALL. Studies that have directly evaluated the learning behaviour in SLA and ALL have found a relationship between a learner's ability to learn a second language and their performance in an ALL experiment (see Ettlinger, Morgan-Short, Faretta-Stutenberg, and Wong (2016) and references therein). Both areas seek to understand the factors that underlie language learning. In SLA experimental studies often consist of a learning phase and a test phase for investigating language learning behaviours of different L2s by speakers of different L1s (Muylle, Bernolet, & Hartsuiker, 2021; Wonnacott, Brown, & Nation, 2017). The selected language material can be real, artificial, or a mixture of the two. Crucially, our approach can be readily applied in SLA research, since the choice of training words often depends mainly on the criteria of structural simplicity, learner-friendliness, and is not controlled with respect to L1 and their number is limited (Sinkeviciute, Brown, Brekelmans, & Wonnacott, 2019).

Third, we want to note that the concept can also be applied to other types of learning (i.e., investigation of syntactic learning and semantic learning, such as Muylle et al., 2021; Snefjella, Lana, & Kuperman, 2020). Unlike in our experiment, in which we assessed L1/AL knowledge with the help of lexicons including information about phonological features and frequencies, one could use suitable input sources or corpora matching the needs of the specific investigation. Let us consider the following two cases: a) Wordlikeliness in phonological alternation learning, and b) measures beyond lexical statistics. First, wordlikeliness could affect the learning of phonological alternations. For instance, if the alternation learning task is to choose one of two options (e.g., choose the plural form given a singular form), or rate a given option as in Albright and Hayes (2003) one could include in the analysis the wordlikeness of each option given their L1 or the training stimuli. Second, one could include measures of patterns of other kinds. For instance, in an experiment that seeks to examine the preference of two syntactic word orders (Subject-Verb-Object, and Object-Verb-Subject), the distribution of these two word orders in the learners' L1

can be estimated as the frequency ratio of the two patterns as appeared in a text corpus. In both cases, these measures (be they lexical or syntactic) can then be captured as covariates during statistical modelling just like in our study. Linguistic L1 and AL knowledge of any kind can be assessed and evaluated in investigations of language learning.

Finally, we want to conclude that considering the effects of L1 and AL, as well as the task, should be incorporated in existing models of language learning (Albright, 2009; Bylinina, Tikhonov, & Garmash, 2021; Hayes & Wilson, 2008; Johnson, Culbertson, Rabagliati, & Smith, 2020; Kakolu Ramarao et al., 2023; Linzen & O'Donnell, 2015; Tang, Kakolu Ramarao, & Baer-Henney, 2022) in order to better simulate human learning behaviour.

5. Conclusions

The present study investigated the influence of L1 and AL lexical statistics during the ALL experiment with adult German and Mandarin speakers. We uncovered rapid learning of an identity pattern, which had originally been found for English speakers by Linzen and Gallagher (2017). Moreover, we found that both the L1 and the AL lexicon contribute to the performance of the learners. While L1 wordlikeness led participants to reject artificial items, AL wordlikeness led participants to accept them. As a consequence, we argue that it is highly important in ALL research to take these measures into account. The effect of rapid learning of the identity constraint in the present paper persisted and could be confirmed for speakers of two very distinct languages despite the impact of L1 and AL statistics.

We show that it is possible to quantify psycholinguistic variables that are known to play a role in language processing in ALL research by incorporating them as covariates in a statistical model. We argue that it is important and possible to take L1 covariates into account, such as neighbourhood density and lexical activation diversity (Bailey & Hahn, 2001; Milin et al., 2017). Also, AL variables can be implemented in a similar way to capture the effect of the trained lexicon. In sum, it is clearly necessary to design the artificial language and control AL lexical statistics with great care, and we have shown how this can be done.

We acknowledge that the ALL paradigm is a great opportunity to investigate learning behaviour under laboratory conditions. It enables us to research the complex questions in language learning in a relatively simple way. ALL research is thus suitable to complement more intricate investigations of children. Underlying learning mechanisms may be similar and thus help us to explain the language learning and acquisition process.

Our arguments are relevant for language learning studies investigating learning behaviour and learning mechanisms of any type. Regardless of whether a study deals with learners (children or adults) that learn any artificial or real L2, the items used in training and the L1 knowledge of the participants can jointly shape learners' behaviour.

Data accessibility statement

The data and code of the analyses can be found on an Open Science Framework repository: https://www.doi.org/10.17605/OSF.IO/M2DTG.

Additional file

The additional files for this article can be found as follows:

- **Appendix.** The appendix consists of four parts A, B, C and D. In part A, we provide task instructions for both the German as well as the Mandarin version of the study. In part B, we provide mathematical details of the Rescorla-Wagner-equations. In part C, we provide all experimental items for frequent and infrequent groups of the German and Mandarin version of the study. In part D, we provide random effects summaries for the German and Mandarin studies. DOI: https://doi.org/10.16995/labphon.6460.s1

Acknowledgements

We thank Steffi Wulff, Ratree Wayland, Fabian Tomaschek and the audience at Manchester Phonology Meeting 2019 and the seminar series "Typical and atypical language acquisition" at the University of Potsdam for their feedback on earlier versions of the study. We thank Stella Qi, Mengjie Zhang, and Charlotte von Kries for their assistance with the recording and processing of the stimuli. We thank Joshua Martin and Christopher Geissler for providing feedback on the writing. We thank Alexander Martin and our anonymous reviewers for their valuable feedback. We thank Natalie Boll-Avetisyan for pointing out relevant ALL studies that examine OCP. Finally, we thank Adam Albright for discussing the analogical model with us. All errors remain ours.

Competing interests

The authors have no competing interests to declare.

Author contributions

Kevin Tang: Conceptualization, Methodology, Data curation, Formal analysis, Writing-Original draft preparation, Visualization, Investigation, Writing-Reviewing and Editing. Dinah Baer-Henney: Conceptualization, Methodology, Data curation, Formal analysis, Writing-Original draft preparation, Visualization, Investigation, Writing-Reviewing and Editing.

References

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41. DOI: https://doi.org/10.1017/S0952675709001705

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161. DOI: https://doi.org/10.1016/S0010-0277(03)00146-X

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, *1*(1), 1–47.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511801686

Baayen, R. H. (2010). A real experiment is a factorial experiment? *Mental Lexicon*, *5*(1), 149–157. DOI: https://doi.org/10.1075/ml.5.1.06baa

Baayen, R. H. (2013). languageR: Data sets and functions with ”analyzing linguistic data: A practical introduction to statistics”. R package version 1.4.1 [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=languageR

Baayen, R. H., Endresen, A., Janda, L. A., Makarova, A., & Nesset, T. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, *37*(3), 253–291. DOI: https://doi.org/10.1007/s11185-013-9118-6

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–482. DOI: https://doi.org/10.1037/a0023851

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Baer-Henney, D. (2015). *Learner's little helper – strength and weakness of the substantive bias in phonological acquisition* (Doctoral dissertation, University of Potsdam). Retrieved from https://publishup.uni-potsdam.de/frontdoor/index/index/docId/8309

Baer-Henney, D., Kügler, F., & van de Vijver, R. (2015). The interaction of languagespecific and universal factors during the acquisition of morphophonemic alternations with exceptions. *Cognitive Science*, *39*(7), 1537–1569. DOI: https://doi.org/10.1111/cogs.12209

Baer-Henney, D., & van de Vijver, R. (2012). On the role of substance, locality and amount of exposure in the acquisition of morphophonemic alternations. *Laboratory Phonology*, *3*(2), 221–249. DOI: https://doi.org/10.1515/lp-2012-0013

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591. DOI: https://doi.org/10.1006/jmla.2000.2756

Bailey, T. M., & Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, *52*(3), 339–362. DOI: https://doi.org/10.1016/j.jml.2004.12.003

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Becker, M., & Levine, J. (2013). Experigen–an online experiment platform. *Available at* http://becker.phonologist.org/experigen.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, New Jersey, US: Wiley. DOI: https://doi.org/10.1002/jae.3950040108

Bennett, R., Tang, K., & Ajsivinac Sian, J. (2018). Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan). *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1), 9. DOI: https://doi.org/10.5334/labphon.100

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177. DOI: https://doi.org/10.1080/00437956.1958.11659661

Blevins, J., Milin, P., & Ramscar, M. (2017). The Zipfian paradigm cell filling problem. In F. Kiefer, J. Blevins, & H. Bartos (Eds.), *Perspectives on morphological structure: Data and analyses* (pp. 139–158). Leiden: Brill. DOI: https://doi.org/10.1163/9789004342934_008

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning Dutch as a second language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747–1760. DOI: https://doi.org/10.1044/2015_JSLHR-L-15-0058

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer. Version 6.0.40.* Retrieved from http://www.praat.org

Boll-Avetisyan, N. (2012). *Phonotactics and its acquisition, representation, and use: an experimental-phonological study* (Unpublished doctoral dissertation). (LOT Dissertation series)

Boll-Avetisyan, N., & Kager, R. (2016). Is speech processing influenced by abstract or detailed phonotactic representations? The case of the obligatory contour principle. *Lingua*, *171*, 74–91. DOI: https://doi.org/10.1016/j.lingua.2015.11.008

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5). DOI: https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. DOI: https://doi.org/10.3758/BRM.41.4.977

Burchfield, L. A., & Bradlow, A. R. (2014). Syllabic reduction in Mandarin and English speech. *The Journal of the Acoustical Society of America*, *135*(6), 270–276. DOI: https://doi.org/10.1121/1.4874357

Bylinina, L., Tikhonov, A., & Garmash, E. (2021). *Old bert, new tricks: Artificial language learning for pre-trained language models.* arXiv. Retrieved from https://arxiv.org/abs/2109.06333. DOI: https://doi.org/10.48550/ARXIV.2109.06333

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One*, *5*(6), e10729. DOI: https://doi.org/10.1371/journal.pone.0010729

Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, *27*(3), 345–392. DOI: https://doi.org/10.1017/S0952675710000199

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77. DOI: https://doi.org/10.1016/s0010-0277(02)00233-0

Chen, F.-Y., Mo, R.-P., Huang, C.-R., & Chen, K.-J. (1992). Reduplication in Mandarin Chinese: Their formation rules, syntactic behavior and ICG representation. In *Proceedings of R.O.C. Computational Linguistics Conference v* (pp. 217–233). Taipei, Taiwan. Retrieved from https://aclanthology.org/O92-1007.pdf

Chen, T.-Y., & Myers, J. (2021). Worldlikeness: a web-based tool for typological psycholinguistic research. *Linguistics Vanguard*, *7*(s1), 20190011. DOI: https://doi.org/10.1515/lingvan-2019-0011

Chuang, Y.-Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2019). On the processing of nonwords in word naming and auditory lexical decision. In S. Calhoun, P. Escudero, T. Marija, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1233–1237). Canberra, Australia: Australasian Speech Science and Technology Association Inc. DOI: https://doi.org/10.31234/osf.io/ekvma

Chuang, Y.-Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension: a computational modeling approach using linear discriminative learning. *Behavioural Research Methods*, *53*(3), 945–976. DOI: https://doi.org/10.3758/s13428-020-01356-w

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, *30*(2), 441–469. DOI: https://doi.org/10.1017/S0305000903005579

Cristià, A., & Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? *Language Learning and Development*, *4*(3), 203–227. DOI: https://doi.org/10.1080/15475440802143109

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82. DOI: https://doi.org/10.1016/j.cognition.2015.02.007

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. DOI: https://doi.org/10.1016/j.cognition.2011.10.017

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121. DOI: https://doi.org/10.1016/S0022-2496(02)00016-0

Davidson, D., & Martin, A. E. (2013). Modeling accuracy as a function of response time with the generalized linear mixed effects model. *Acta Psychologica*, *144*(1), 83–96. DOI: https://doi.org/10.1016/j.actpsy.2013.04.016

de Chene, B. (2014). Probability matching versus probability maximization in morphophonology: The case of Korean noun inflection. *Theoretical and Applied Linguistics at Kobe Shoin*, *17*, 1–13.

DeMille, M. M. C., Tang, K., Mehta, C. M., Geissler, C., Malins, J. G., Powers, N. R., … Gruen, J. R. (2018). Worldwide distribution of the *DCDC2* READ1 regulatory element and its relationship with phoneme variation across languages. *Proceedings of the National Academy of Sciences*. DOI: https://doi.org/10.1073/pnas.1710472115

Denisowski, P. (1997). *Cedict: Chinese-English dictionary*. Retrieved 2014-07-01, from http://www.mdbg.net/chindict/chindict.php?page=cc-cedict

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford: Oxford University Press. DOI: https://doi.org/10.1017/S0952675701004195

Durvasula, K., & Liter, A. (2020). There is a simplicity bias when generalising from ambiguous data. *Phonology*, *37*(2), 177–213. DOI: https://doi.org/10.1017/S0952675720000093

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). Wordgen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 488–499. DOI: https://doi.org/10.3758/BF03195595

Eden, S. E. (2018). *Measuring phonological distance between languages* (Doctoral dissertation, UCL (University College London)). Retrieved from https://discovery.ucl.ac.uk/id/eprint/10058348/1/Eden_10058348_thesis.pdf

Edwards, J. G. H., & Zampini, M. L. (2010). *Phonology and second language acquisition*. Philadelphia: John Benjamins Publishing. DOI: https://doi.org/10.1111/j.1540-4781.2009.00994.x

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M., & Wong, P. C. (2016). The relationship between artificial and second language learning. *Cognitive Science*, *40*(4), 822–847. DOI: https://doi.org/10.1111/cogs.12257

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180. DOI: https://doi.org/10.1016/j.jml.2016.06.002

Féry, C. (1998). German word stress in optimality theory. *Journal of Comparative Germanic Linguistics*, *2*(2), 101–142. DOI: https://doi.org/10.1023/A:1009883701003

Finley, S. (2011). The privileged status of locality in consonant harmony. *Journal of Memory and Language*, *65*(1), 74–83. DOI: https://doi.org/10.1016/j.jml.2011.02.006

Finley, S. (2012). Typological asymmetries in round vowel harmony: Support from artificial grammar learning. *Language and Cognitive Processes*, *27*(10), 1550–1562. DOI: https://doi.org/10.1080/01690965.2012.660168

Finley, S. (2017). Learning metathesis: Evidence for syllable structure constraints. *Journal of Memory and Language*, *92*, 142–157. DOI: https://doi.org/10.1016/j.jml.2016.06.005

Finley, S. (2022). Generalization to novel consonants: Place versus voice. *Journal of Psycholinguistic Research*, *51*(6), 1–27. DOI: https://doi.org/10.1007/s10936-022-09897-1

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*(3), 423–437. DOI: https://doi.org/10.1016/j.jml.2009.05.002

Finley, S., & Badecker, W. (2012). Learning biases for vowel height harmony. *Journal of Cognitive Science*, *13*(3), 287–327. DOI: https://doi.org/10.17791/jcs.2012.13.3.287

Frisch, S. (1996). *Similarity and frequency in phonology* (Doctoral dissertation, Northwestern University). Retrieved from https://rucore.libraries.rutgers.edu/rutgers-lib/37658/PDF/1/play/

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? it all depends on the nonwords. *Memory & Cognition*, *23*(1), 83–94. DOI: https://doi.org/10.3758/BF03210559

Glewwe, E. (2019). *Bias in phonotactic learning: Experimental studies of phonotactic implicationals* (Doctoral dissertation, UCLA). Retrieved from https://escholarship.org/content/qt4456s1j0/qt4456s1j0.pdf

Goldsmith, J. (1976). *Autosegmental Phonology* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from http://oastats.mit.edu/bitstream/handle/1721.1/16388/03188555-MIT.pdf?sequence=1&isAllowed=y

Gorman, K., & Johnson, D. E. (2013). Quantitative analysis. In R. Bayley, R. Cameron, & C. Lucas (Eds.), *The Oxford Handbook of Sociolinguistics* (p. 214–240). Oxford, UK: Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780199744084.013.0011

Graff, P., & Jaeger, T. (2009). Locality and feature specificity in OCP effects: Evidence from Aymara, Dutch, and Javanese. In *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (Vol. 45, pp. 127–141). Chicago: Chicago Linguistic Society. Retrieved from https://www.researchgate.net/profile/T-Florian-Jaeger/publication/229079018_Locality_and_Feature_Specificity_in_OCP_Effects_Evidence_from_Aymara_Dutch_and_Javanese/links/0f317536eeb015807a000000/Locality-and-Feature-Specificity-in-OCP-Effects-Evidence-from-Aymara-Dutch-and-Javanese.pdf

Greenwood, A. (2016). *An experimental investigation of phonetic naturalness* (Doctoral dissertation, UC Santa Cruz). Retrieved from https://escholarship.org/content/qt94x407sb/qt94x407sb.pdf

Günther, F., Smolka, E., & Marelli, M. (2019). 'Understanding' differs between English and German: Capturing systematic language differences of complex words. *Cortex*, *116*, 168–175. DOI: https://doi.org/10.1016/j.cortex.2018.09.007

Harris, J., Neasom, N., & Tang, K. (In prep). Phonotactics with [awt] rules: the learnability of a simple, unnatural pattern in English.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. DOI: https://doi.org/10.1162/ling.2008.39.3.379

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150. DOI: https://doi.org/10.3389/fnins.2014.00150

Hendrix, P. (2016). *Experimental explorations of a discrimination learning approach to language processing* (Doctoral dissertation, University of Tübingen). Retrieved from https://ub01.uni-tuebingen.de/xmlui/bitstream/handle/10900/67914/thesisPeterHendrix.pdf?sequence=2&isAllowed=y

Howell, P., Tang, K., Tuomainen, O., Chan, S. K., Beltran, K., Mirawdeli, A., & Harris, J. (2017). Identification of fluency and word-finding difficulty in samples of children with diverse language backgrounds. *International Journal of Language & Communication Disorders*, *52*(5), 595–611. DOI: https://doi.org/10.1111/1460-6984.12305

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195. DOI: https://doi.org/10.1207/s15473341lld0102_3

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, *122*(5), 2842–2854. DOI: https://doi.org/10.1121/1.2783198

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877. DOI: https://doi.org/10.1121/1.3148196

Johnson, T., Culbertson, J., Rabagliati, H., & Smith, K. (2020, Mar). *Assessing integrative complexity as a predictor of morphological learning using neural networks and artificial language learning*. PsyArXiv. Retrieved from psyarxiv.com/yngw9. DOI: https://doi.org/10.31234/osf.io/yngw9

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (2nd edition)*. Upper Saddle River, New Jersey: Prentice Hall.

Kakolu Ramarao, A., Tang, K., & Baer-Henney, D. (2023). *Can Neural Networks learn humanlike behavior when inflecting verbs? The case of Spanish.* (Presentation at the Five-Minute linguist competition, *Linguistic Society of America 97th Annual Meeting 2023*, Denver, CO, USA)

Kakolu Ramarao, A., Zinova, Y., Tang, K., & van de Vijver, R. (2022, July). HeiMorph at SIGMORPHON 2022 shared task on morphological acquisition trajectories. In *Proceedings of the 19th sigmorphon workshop on computational research in phonetics, phonology, and morphology* (pp. 236–239). Seattle, Washington: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.sigmorphon-1.24

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. DOI: https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. DOI: https://doi.org/10.3758/BRM.42.3.643

Kroll, J. F., & De Groot, A. M. (2009). *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford University Press.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lin, I. (2016). Tone Sequences in Lexical Processing of Beijing Mandarin. *The Journal of the Acoustical Society of America*, *140*(4), 3224. DOI: https://doi.org/10.1121/1.4970181

Linzen, T., & Gallagher, G. (2017). Rapid generalization in phonotactic learning. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1), 1–32. DOI: https://doi.org/10.5334/labphon.44

Linzen, T., & O'Donnell, T. (2015). A model of rapid phonotactic generalization. In *Proceedings of the Empirical Methods in Natural Language processing EMNLP 2015* (pp. 1126–1131). Retrieved from https://aclanthology.org/D15-1134. DOI: https://doi.org/10.18653/v1/D15-1134

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. DOI: https://doi.org/10.1097/00003446-199802000-00001

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, *52*(3), 436–459. DOI: https://doi.org/10.1016/j.jml.2005.01.013

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. DOI: https://doi.org/10.3758/s13428-016-0809-y

Martin, A., & White, J. (2021). Vowel harmony and disharmony are not equivalent in learning. *Linguistic Inquiry*, *52*(1), 227–239. DOI: https://doi.org/10.1162/ling_a_00375

McMullin, K., & Hansson, G. O. (2016). Long-distance phonotactics as tier-based strictly 2-local languages. In A. Albright & M. A. Fullwood (Eds.), *Proceedings of the Annual Meetings on Phonology* (Vol. 2). Washington DC: PKP. Retrieved from https://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/3750/3468. DOI: https://doi.org/10.3765/amp.v2i0.3750

Melloni, C., & Basciano, B. (2018). Reduplication across boundaries: The case of Mandarin. In O. Bonami, G. Boyé, G. Dal, H. Giraudo, & F. Namer (Eds.), *The lexeme in descriptive and theoretical morphology* (p. 325–362). Berlin: Language Science.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS One*, *12*(2), e0171935. DOI: https://doi.org/10.1371/journal.pone.0171935

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*, 338–352. DOI: https://doi.org/10.1121/1.1907526

Moran, S., McCloy, D., & Wright, R. (Eds.) (2014). *Phoible online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://phoible.org/Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127. DOI: https://doi.org/10.1017/S0952675708001413

Moreton, E. (2012). Inter-and intra-dimensional dependencies in implicit phonotactic learning. *Journal of Memory and Language*, *67*(1), 165–183. DOI: https://doi.org/10.1016/j.jml.2011.12.003

Moreton, E., & Pater, J. (2012a). Structure and substance in artificial-phonology learning. Part II: Substance. *Language and Linguistics Compass*, *6*(11), 702–718. DOI: https://doi.org/10.1002/lnc3.366

Moreton, E., & Pater, J. (2012b). Structure and substance in artificial-phonology learning, Part I: Structure. *Language and Linguistics Compass*, *6*(11), 686–701. DOI: https://doi.org/10.1002/lnc3.363

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1033–1047. DOI: https://doi.org/10.1044/1092-4388(2005/072)

Muylle, M., Bernolet, S., & Hartsuiker, R. J. (2021). The development of shared syntactic representations in late l2-learners: Evidence from structural priming in an artificial language. *Journal of Memory and Language*, *119*, 104233. DOI: https://doi.org/10.1016/j.jml.2021.104233

Myers, S., & Padgett, J. (2014). Domain generalisation in artificial language learning. *Phonology*, *31*(3), 399–433. DOI: https://doi.org/10.1017/S0952675714000207

Nevins, A., Rodrigues, C., & Tang, K. (2015, March). The rise and fall of the L-shaped morphome: diachronic and experimental studies. *Probus: International Journal of Latin and Romance Linguistics*, *27*(1), 101–155. DOI: https://doi.org/10.1515/probus-2015-0002

Nicenboim, B., Roettger, T. B., & Vasishth, S. (2018). Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics*, *70*, 39–55. DOI: https://doi.org/10.1016/j.wocn.2018.06.001

Nieder, J., & Tang, K. (2023). *A corpus study of naturalistic misperceptions in German sung speech.* (Phonetics and Phonology in Europe 2023, Radboud University in Nijmegen, the Netherlands)

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., … others (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468. DOI: https://doi.org/10.7554/eLife.33468

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081. DOI: https://doi.org/10.1016/j.cognition.2019.104081

Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. DOI: https://doi.org/10.1037/0096-3445.115.1.39

Onishi, K. H., Chambers, K. E., & Fisher, C. (2002). Learning phonotactic constraints from brief auditory experience. *Cognition*, *83*(1), B13–B23. DOI: https://doi.org/10.1016/S0010-0277(01)00165-2

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268–284. DOI: https://doi.org/10.1016/j.cognition.2012.10.008

Pater, J., & Tessier, A.-M. (2003). Phonotactic knowledge and the acquisition of alternations. In M.-J. Solé, R. Daniel, & J. Romero (Eds.), *Proceedings of the 15th International Congress on Phonetic Sciences* (Vol. 1180, pp. 1177–1180). Barcelona, Spain. Retrieved from http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1177.pdf

Pham, H., & Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience*, *30*(9), 1077–1095. DOI: https://doi.org/10.1080/23273798.2015.1054844

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*(2), 123–138. DOI: https://doi.org/10.1037/0096-3445.125.2.123

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton.

Roettger, T. B., & Baer-Henney, D. (2019). Toward a replication culture: Speech production research in the classroom. *Phonological Data and Analysis*, *1*(4), 1–23. DOI: https://doi.org/10.3765/pda.v1art4.13

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114. DOI: https://doi.org/10.1016/j.cogpsych.2017.02.004

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056. DOI: https://doi.org/10.1016/j.cognition.2019.104056

Schmitz, D., Plag, I., Baer-Henney, D., & Stein, S. (2021). Durational differences of wordfinal /s/ emerge from the lexicon: Modelling morpho-phonetic effects in pseudowords with linear discriminative learning. *Frontiers in Psychology*, *12*. DOI: https://doi.org/10.3389/fpsyg.2021.680889

Seidl, A., & Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, *1*(3–4), 289–316. DOI: https://doi.org/10.1080/15475441.2005.9671950

Shaoul, C., Bitschau, S., Schilling, N., Arppe, A., Hendrix, P., Milin, P., & Baayen, R. H. (2015). ndl2: Naive discriminative learning [Computer software manual]. (R package version 0.1.0.9002, development version available upon request)

Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, *41*(4), 795–820. DOI: https://doi.org/10.1017/S0272263119000263

Skoruppa, K. (2019). Noun and verb learning in an artificial language in mono- and multilingual children: A multilingual verb learning advantage in German-learning first grade school children. *Travaux Neuchâtelois de Linguistique*, *71*, 109–123. DOI: https://doi.org/10.26034/tranel.2019.2994

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160051. DOI: https://doi.org/10.1098/rstb.2016.0051

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449. DOI: https://doi.org/10.1016/j.cognition.2010.06.004

Snefjella, B., Lana, N., & Kuperman, V. (2020). How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language*, *115*, 104171. DOI: https://doi.org/10.1016/j.jml.2020.104171

Stack, C. M. H., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. DOI: https://doi.org/10.3758/s13421-018-0808-6

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192. DOI: https://doi.org/10.1044/1092-4388(2006/085)

Surendran, D., & Niyogi, P. (2003). *Measuring the usefulness (functional load) of phonological contrasts* (Tech. Rep.). Chicago: Department of Computer Science, University of Chicago. Retrieved from https://newtraell.cs.uchicago.edu/research/publications/techreports/TR-2003-12 (Technical Report TR-2003)

Tang, K. (2012). A 61 million word corpus of Brazilian Portuguese film subtitles as a resource for linguistic research. *UCL Working Papers in Linguistics*, *24*, 208–214.

Tang, K. (2015). *Naturalistic speech misperception* (Unpublished doctoral dissertation). University College London.

Tang, K., & Akkuş, F. (2022, Jul). *Identity avoidance in turkish partial reduplication: Feature specificity and locality*. PsyArXiv. Retrieved from psyarxiv.com/vbn6p. DOI: https://doi.org/10.31234/osf.io/vbn6p

Tang, K., Chang, C. B., Green, S., Bao, K. X., Hindley, M., Kim, Y. S., & Nevins, A. (2022). Intoxication and pitch control in tonal and non-tonal language speakers. *JASA Express Letters*, *2*(6), 065202. DOI: https://doi.org/10.1121/10.0011572

Tang, K., & de Chene, B. (2014). *A new corpus of colloquial Korean and its applications.* (Poster presented at *the 14th Conference on Laboratory Phonology*, Tachikawa, Tokyo, Japan.)

Tang, K., DeMille, M. M. C., Frijters, J. C., & Gruen, J. R. (2020). *DCDC2* READ1 regulatory element: how temporal processing differences may shape language. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1928), 20192712. DOI: https://doi.org/10.1098/rspb.2019.2712

Tang, K., Kakolu Ramarao, A., & Baer-Henney, D. (2022). *Modeling irregular morphological patterns with recurrent neural network: the case of the L-shaped morphome.* (Poster presented at *13th Mediterranean Morphology Meeting*, University of the Aegean, Greece.)

Tang, K., & Nevins, A. (2013). Quantifying the diachronic productivity of irregular verbal patterns in Romance. *UCL Working Papers in Linguistics*, *25*, 289–308.

Tang, K., & Nevins, A. (2014). Measuring segmental and lexical trends in a corpus of naturalistic speech. In H.-L. Huang, E. Poole, & A. Rysling (Eds.), *Proceedings of the 43rd meeting of the North East Linguistic Society* (Vol. 2, pp. 153–166). GLSA (Graduate Linguistics Student Association).

Tang, K., & Shaw, J. A. (2021). Prosody leaks into the memories of words. *Cognition*, *210*, 104601. DOI: https://doi.org/10.1016/j.cognition.2021.104601

van de Vijver, R., & Baer-Henney, D. (2014). Developing biases. *Frontiers in Psychology*, *5*(634), 1–7. DOI: https://doi.org/10.3389/fpsyg.2014.00634

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. DOI: https://doi.org/10.1080/17470218.2013.850521

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item learning. *Journal of Memory and Language*, *119*, 104231. DOI: https://doi.org/10.1016/j.jml.2021.104231

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., … van De Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial

language learning. In S. Hucklebridge & M. Nelson (Eds.), *Nels 48: Proceedings of the Forty-Eighth Annual Meeting of the North East Linguistic Society* (Vol. 3, p. 207–220). Amherst, MA, USA: GLSA. Retrieved from https://dspace.library.uu.nl/bitstream/handle/1874/421973/White_Kager_Linzen_Markopoulos_Martin_Nevins_Peperkamp_Polgardi_Topintzi_van_de_Vijver_2018_NELS_48.pdf?sequence=1

White, J., & Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition*, *133*(1), 85–90. DOI: https://doi.org/10.1016/j.cognition.2014.05.020

Wiese, R. (2000). *The Phonology of German*. Oxford: Oxford University Press.

Wilson, C. (2003). Experimental investigation of phonological naturalness. In G. Garding & M. Tsujimura (Eds.), *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (Vol. 22, pp. 533–546). Sommerville, MA: Cascadilla Press. Retrieved from https://linguistics.ucla.edu/people/wilson/Wilson2003.pdf

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945–982. DOI: https://doi.org/10.1207/s15516709cog0000_89

Wissmann, M., Toutenburg, H., & Shalabh. (2007). *Role of categorical variables in multicollinearity in the linear regression* (Tech. Rep. No. 008). Munich, Germany: Department of Statistics, University of Munich. Retrieved from https://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, *95*, 36–48. DOI: https://doi.org/10.1016/j.jml.2017.01.005

Yin, S. H., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. *Cognition*, *179*, 89–101. DOI: https://doi.org/10.1016/j.cognition.2018.05.023

Zhou, X., & Marslen-Wilson, W. (1994). Words, morphemes and syllables in the Chinese mental lexicon. *Language and Cognitive Processes*, *9*(3), 393–422. DOI: https://doi.org/10.1080/01690960802174514