

Phonetically natural rules benefit from a learning bias: a re-examination of vowel harmony versus disharmony

Alexander Martin^{1,2} & Sharon Peperkamp²

¹*Centre for Language Evolution
University of Edinburgh
3 Charles Street, Edinburgh EH8 9AD, United Kingdom*

²*Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)
Département d'Études Cognitives, École Normale Supérieure – PSL University
29, rue d'Ulm, 75005 Paris, France*

Abstract: Phonological rules tend to be phonetically ‘natural’: they reflect constraints on speech production and perception. Substance-based phonological theories predict that a preference for phonetically natural rules is encoded in synchronic grammars and translates into learning biases. Some previous work has shown evidence for such biases, but methodological concerns with those studies mean that the question warrants further investigation. We revisit this issue by focussing on the learning of palatal vowel harmony (phonetically natural) compared to disharmony (phonetically unnatural). In addition, we investigate the role of memory consolidation during sleep on rule learning. We use an artificial language learning paradigm with two test phases separated by twelve hours. We observe a robust effect of phonetic naturalness: vowel harmony is learned better than vowel disharmony. For both rules, performance remains stable after 12 hours, regardless of the presence vs absence of sleep.

Keywords: artificial language learning, learning bias, phonetic naturalness, memory consolidation, sleep, vowel harmony

1 Introduction

Sound patterns tend to be phonetically ‘natural’: they reflect constraints on speech production and perception. For instance, many phonological rules, such as consonant assimilation or vowel harmony, increase ease of articulation; they mirror low-level gradient phonetic effects due to automatic processes such as coarticulation and gestural overlap. It has long been observed that cross-linguistically, phonetically natural rules are much more prevalent than unnatural ones (Hooper, 1976). One explanation for such typological asymmetries concerns the existence of an individual learning bias. In various phonological theories, typological asymmetries are reflected by phonetically motivated biases in the speaker’s mind (Archangeli & Pulleyblank, 1994; Donegan & Stampe, 1979; Hayes & Steriade, 2004). These grammatical biases could induce learning biases, such that phonetically natural patterns are easier to learn and hence have an advantage in transmission across generations (Schane, Tranel, & Lane, 1974; Wilson, 2006). Note that while some theories posit transmission as the sole locus of the evolution of natural patterns over unnatural ones (Blevins, 2004; Ohala, 1993), the learning bias hypothesis is not strictly speaking incompatible with them (Beguš, 2018; Moreton, 2008). The learning bias hypothesis has repeatedly been put to the test, exploring the idea of a learning bias for many different typological regularities. Some evidence has been found in

favour of a learning bias, though results from such studies make it difficult to disentangle phonetic naturalness from complexity.

Indeed, there is ample evidence for a learning bias favouring simpler patterns over more complex ones (for a review, see Moreton & Pater, 2012). For example, White (2014) showed that participants more easily learn simple phonological alternations than saltatory ones (which are rare, but attested). Saltatory alternations involve an alternation where an intermediate segment must be “jumped over” (e.g., a rule that says an underlying segment /p/ surfaces as [v] between vowels, jumping over [b]). Such alternations, though, affect multiple phonological features, making them featurally more complex than an alternation affecting only one feature (e.g., intervocalic voicing of /p/ to [b] or spirantisation of /b/ to [v]).¹ In a similar vein, Skoruppa, Lambrechts, & Peperkamp (2011) showed quicker and better learning of single-feature alternations than of multi-feature alternations. Likewise, Peperkamp, Skoruppa, & Dupoux (2006) found that a rule applying to a typologically attested natural class is learnt better than one applying to an arbitrary group of sounds.

Still, some evidence of a learning bias for some phonological patterns over others that does not seem to involve complexity does exist. For example, Schane, Tranel, & Lane (1974) found that a typologically attested rule of word-final consonant epenthesis before a vowel (akin to the rule of French liaison) is learnt faster compared to a typologically unattested one of word-final consonant *deletion* before a vowel. Wilson (2006) demonstrated asymmetrical generalisation of a newly learnt velar palatalisation rule, with participants generalising the rule from before mid to before high vowels, but not vice versa. This follows the typological, phonetically-grounded fact that languages that palatalise velar stops before mid vowels also do so before high vowels, whereas the inverse is not necessarily true. It should be noted, though, that this study contained other experimental conditions that did not yield learning asymmetries that are predicted by typological facts. Another example of evidence for bias towards phonetic naturalness comes from Myers and Padgett (2014). They showed not only that the phonetically natural rule of final devoicing is learnt better than the phonetically unnatural rule of final voicing, but also that the former is better generalised to utterance-medial position than the latter, following typological facts. The participants in that study, however, were native English speakers and English is known to be subject to phonetic devoicing (e.g., Docherty, 1992), which might have influenced the participants’ performance. Beyond segmental patterns, there is evidence of a learning bias for rules targeting suprasegmental phenomena as well. A study with English- and French-speaking participants considered learning of a rule where low vowels attracted stress (phonetically natural) compared to a rule where high vowels attracted stress (phonetically

¹ White claimed that his full results could not be simply reduced to a question of complexity, since participants in another condition showed asymmetrical preferences regarding single-feature alternations. Specifically, they preferred to change voiced stops to voiceless stops (e.g., /b/ > [p]) rather than voiceless fricatives to voiceless stops (e.g., /f/ > [p]). This asymmetry is consistent with other work showing that single-feature differences are not all perceived as equally distinct, and that place and manner of articulation may have preferential status in perception compared to voicing (Martin & Peperkamp, 2017).

unnatural) (Carpenter, 2010). In that study, the natural rule was overall learned better than the unnatural one. More recently, the same author likewise taught English- and French-speaking participants a stress assignment rule whereby stress was attracted to the leftmost heavy syllable (natural) or a similar rule whereby stress was attracted to the leftmost *light* syllable (unnatural). The results showed better learning of the former than of the latter (Carpenter, 2016).²

In this article, we focus on vowel harmony. This common phonological phenomenon involves co-occurrence restrictions on vowels, such that all of the vowels within a certain domain (typically the word) must share one or more phonological features. It often presents itself in the form of morphophonological alternations. Hungarian, for example, has a restriction on the backness of vowels within a word, such that most suffixes of the language have two allomorphs: one containing a back vowel, and one containing a front vowel. The data in (1) demonstrates this restriction, where (1a) contains only back vowels and (1b) contains only front vowels, though both contain the same dative suffix.

- (1) a. [bɒra:t-nɒk]
friend-DAT
b. [ɛmber-nɛk]
person-DAT

Some experimental work concerning learning bias for vowel harmony has compared it to alternations affecting arbitrary groups of sounds (Baer-Henney, Kügler, & van de Vijver, 2014; Pycha, Nowak, Shin, & Shosted, 2003). These studies have shown that vowel harmony is learnt better than the arbitrary alternations. Similar studies have found better learning of vowel harmony than a vowel-consonant dependency (e.g., where the roundness of a suffix vowel depends on the voicing of a stem consonant (Baer-Henney & van de Vijver, 2012)). The unnatural rules in all of those studies, though, were formally more complex than the vowel harmony rule, in that they concerned multiple features.

There are, though, a number of typological tendencies concerning vowel harmony that do not confound phonetic naturalness with complexity, many of which have been the subject of previous research looking at learning bias. For instance, Finley & Badecker (2008) considered the fact that vowel harmony is always directional in nature (i.e., vowel features spread either from left-to-right or from right-to-left, in line with coarticulatory patterns), and there are no rules in the typology based on a majority count, whereby the feature that occurs most often in the word spreads to the vowels

² Given that in English but not in French stress is sensitive to syllable weight, the results for the French participants are the important ones, and it should be noted that the observed naturalness effect in this group is significant only in a one-tailed t-test; as mentioned by the author, the validity of the use of such a test is not agreed upon.

without that feature. Accordingly, when exposed to input that is compatible with both a directionality-based and a majority-based harmony rule, participants overwhelmingly infer the former. Similar results were reported in an additional study (Finley & Badecker, 2009b). The same authors also considered the fact that vowel harmony causes agreement of subsegmental features (e.g., [BACK]), and showed that participants do indeed base generalisation on such features rather than on individual segments (Finley & Badecker, 2009a), though insofar as segments are themselves composed of multiple features, it is unclear if this result truly disentangles complexity (one vs many features) from naturalness (features vs segments). Some types of harmony show asymmetries in the typology. In rounding harmony systems, mid vowels may trigger harmony for all vowel types while high vowels do not, but the converse is not true. That is, if a language has a rounding harmony system with unrestricted high vowel triggers, it also tends to have unrestricted mid vowel triggers. This asymmetry is hypothesised to enhance the perceptual salience of harmony patterns and boost the perception of the non-high vowels which tend to have weaker phonetic cues to rounding. (Thus, having all the vowels of the word agree in this feature means the listener is more likely to correctly identify the presence of rounding.) Finley (2012) showed better learning of rounding harmony patterns that were triggered by mid vowels than those that were triggered by high vowels, in line with phonetic naturalness and the typology. Taking this one step further, and in a similar vein to Wilson (2006), Kimper (2016) looked at extrapolation in cases of rounding harmony. He taught listeners rounding harmony patterns and showed that learners actively extrapolated the harmony pattern from high to mid vowels, but not from mid to high vowels, again in line with the typology and the phonetic grounding.

In this article, we focus on a further typological tendency concerning vowel harmony: The logically equivalent rule of vowel disharmony, whereby suffixes must *disagree* along some phonological dimension with vowels in the root, is virtually unattested cross-linguistically. Vowel harmony has been proposed to be born out of vowel-to-vowel coarticulation (Ohala, 1994)³, which could explain why disharmony patterns would be less likely to emerge (i.e., they lack a clear phonetic precursor and are thus ‘unnatural’). But if speakers possess phonetic knowledge, could this not bias them towards the learning of a phonetically natural harmony pattern over an unnatural disharmony one? Then, in addition to unnatural rules like vowel disharmony being less likely to arise, they would be less likely to survive repeated transmission, as learners are biased against them (or rather, towards their natural counterparts). This would heavily disadvantage unnatural rules over time and could explain in part why they are so rare in the typology.

Looking for a naturalness bias in the case of harmony versus disharmony presents a clear design advantage compared to all the studies concerning such a bias reviewed above: the test items in

³ The specific ‘natural’ basis of vowel harmony (i.e., whether it is based in a production or perception constraint) is not immediately relevant to our research question. Crucially, vowel disharmony lacks a clear phonetic precursor and is thus ‘unnatural’ according to the simple definition of phonetic naturalness that we consider in this article.

the two conditions are exactly the same. Hence, any observed effect between conditions must be due to the experimental manipulation without being possibly confounded by properties of the items. It should be noted also that a learning bias favouring harmony over disharmony is difficult to explain simply in terms of complexity, since they both depend on a single abstract feature. Depending on the formalisation, it could be argued that disharmony is more complex than harmony because its description requires the use of a negative operator. Work from developmental psychology has shown that the concept ‘different’ may be harder to learn than the concept ‘same’, since the presence of negation makes ‘different’ more complex than ‘same’ (Hochmann, Carey, & Mehler, 2018). From a constraint-based perspective, though, it is the harmony, not the disharmony, pattern that is formalised with the negative operator (e.g., harmony: $*[\alpha F][-\alpha F]$; disharmony: $*[\alpha F][\alpha F]$). We thus remain agnostic with regards to the impact of the negation on the pattern and consider vowel harmony and vowel disharmony to be on a par, as they both depend on a single feature. These single-feature patterns make a good test case for the naturalness learning bias hypothesis.

Two previous studies have failed to show evidence for a learning bias in favour of vowel harmony compared to vowel disharmony. In one of them, American English listeners were exposed in an artificial language learning paradigm to singular/plural alternations that were either harmonic or disharmonic. The numeric pattern of results suggested an advantage for the harmony rule, but this difference was not statistically significant (Pycha et al., 2003). As noted by the authors themselves, with a sample size of only 10 participants per group this study might have been underpowered. In another study with 30 participants per group, though, similar results were obtained: French listeners who were exposed to short stories in a novel accent of their native language which followed a systematic rule of harmony or disharmony showed equal learning of both rules (Skoruppa & Peperkamp, 2011). It is worth noting that both studies also contained a condition in which the participants were exposed to harmony for some vowels and disharmony for others. Interestingly, and in line with the work cited above, in both studies, performance on this more complex rule was significantly worse than on both the harmony and the disharmony rule. This suggests that if there is a bias favouring natural rules at all, it is not as strong as the one favouring structurally less complex rules (Moreton & Pater, 2012b).

We re-examine the question of a learning bias favouring harmony over disharmony. Testing two rules that are equivalent with regards to their complexity, we address the power issues highlighted in previous work. We additionally investigate a novel factor that might influence the learning of phonological rules, namely sleep. Sleep is known to enhance the learning process by way of memory consolidation (Walker & Stickgold, 2004). Newly acquired knowledge consolidates overnight, yielding improved performance the following day. In the domain of language, sleep-dependent memory consolidation has been shown in adults for perceptual adaptation to synthetic speech (Fenn, Nusbaum, & Margoliash, 2003), as well as the learning of non-native sounds (Earle & Myers, 2015), phonotactic constraints (Gaskell et al., 2014), words (Davis, Di Betta, Macdonald, & Gaskell, 2009;

Dumay & Gaskell, 2007; Havas et al., 2018), and morphosyntactic rules (Batterink, Oudiette, Reber, & Paller, 2014). Here, we examine whether sleep enhances the learning of phonological rules, and if so, whether it does so differentially for phonetically natural versus unnatural rules. In particular, if there is only a small learning bias favouring phonetically natural rules but an additional effect of sleep, with more consolidation for these rules compared to their unnatural counterparts, this would add to the evidence of a role for learning biases in shaping the typology.

We use an artificial language learning experiment administered in two sessions (test and retest) separated by twelve hours, either with or without an intervening period of sleep. Like in Pycha et al. (2003), our test case is palatal harmony, a rule whereby vowels within the domain of the word must share the same value along the front/back dimension. This long-distance dependency between vowels is well attested in the typology, but its converse is not. That is, we know of only one language that has been reported to have a possible case of productive palatal *disharmony*, i.e., Ainu (Krämer, 1999).

Given the logistical difficulties of testing participants in the morning and the evening with a 12-hour interval in the lab, we opted for an online experimental setup, allowing participants to take part in the study from home. Specifically, we recruited American English-speaking participants on Mechanical Turk⁴ and tested them on stimuli produced by a native speaker of French from a previous, unpublished, study. Mechanical Turk has been successfully exploited previously using artificial language learning to test learning biases for syntactic typological universals (Culbertson & Adger, 2014; Martin, Abels, Adger, & Culbertson, 2019). Additionally, Steele, Denby, Chan, & Goldrick (2015) used this platform to compare implicit phonotactic learning with native vs non-native stimuli; they found that American English participants were able to learn implicit rules with either English or French stimuli similarly well. Our setup has two main advantages: First, as Steele et al. pointed out, online testing allows for large sample sizes to be recruited quickly and requires relatively few resources on the part of the experimenter (although for the present case with a precisely timed second test session these advantages were less clear, as we will see in section 2.1.3). Second, the use of non-native stimuli reduces the likelihood that participants rely on a metalinguistic strategy to perform the task. As the sounds of one's native language have fixed mappings to orthographic symbols, participants in artificial language learning experiments might encode the stimuli in terms of orthographic rather than phonological categories. Non-native stimuli encourage more phonetic listening, as naïve listeners are less likely to have fixed mappings between non-native stimuli and graphemes.

⁴ Crowd-sourcing participants for experimental research has become more and more common, including in the domain of speech processing (for a detailed discussion, see Eszékéni, Levow, Meng, Parent, & Suendermann, 2013).

2 Experiment

2.1 Methods

2.1.1 Stimuli

Ninety-six CVCV items containing French phonemes, each composed of two different consonants and two different vowels, were used. Half of the items contained two front vowels, drawn from the set /i, e, \tilde{e} /, while the other half contained two back vowels, drawn from the set /u, o, \tilde{o} ⁵; consonants were drawn from the set /b, d, g, p, t, k, v, z, ζ , f, s, \int , n, \varkappa /. Each of the 12 possible vowel combinations (/i-e/, /e-i/, /i- \tilde{e} /, / \tilde{e} -i/, /e- \tilde{e} /, / \tilde{e} -e/, and likewise for the back vowels) occurred equally often (i.e., in eight items). The 96 stimuli were divided into three sets A, B, and C, each containing 16 front stems and 16 back stems. These sets were used variously as exposure or test sets as detailed below.

For each of these items, which were to be used as ‘singulars’, two ‘plurals’ were created, one containing a front vowel, and the other a back vowel: CVCV+/t ϵ l/ and CVCV+/t \omicron l/, respectively. Thus, half of the plural forms were harmonic (i.e., CVCV with front vowels + /t ϵ l/, for instance /pegit ϵ l/, or CVCV with back vowels + /t \omicron l/, for instance /g \ddot{o} dut \omicron l/), and half of them were disharmonic (i.e., CVCV with front vowels + /t \omicron l/, for instance /pegit \omicron l/, or CVCV with back vowels + /t ϵ l/, for instance /g \ddot{o} dut ϵ l/).

All items (singular and plural) were recorded in a soundproof booth by a female native speaker of French using an M-Audio Micro Track II digital recorder and an M-Audio DMP3 pre-amplifier in 16-bit mono at a sampling rate of 44.1 kHz.

2.1.2 Procedure

The experiment was run online from a server in our lab. Participants were therefore not present in the lab and all interfacing with them was done via email. Upon logging into our website for the first time, they were forewarned that there were two parts in the experiment and that they would need to complete the second part between 12 and 13 hours after completing the first part; they were asked only to accept to participate in the experiment if they were sure they would be able to do so. Those who accepted were then asked to provide their email address so that they could be contacted when it was time to come back for the second session.

During the first session, participants received instructions regarding the exposure phase. They were told that they would hear words from an invented language, and that words would be presented in their singular and plural forms. The language was said to have two forms of the plural suffix: ‘tel’ and ‘tol’.

Participants were first exposed to two repetitions of one of the sets A, B, or C (randomly assigned) of 32 unique stems. During exposure, a stem was played, followed after 500 ms by its plural

⁵ Note that all of the front vowels were also unrounded while the back vowels were rounded. Though this confounds roundness and palatal harmony, it serves only to further separate the two groups of vowels.

form (either harmonic or disharmonic, depending on the condition, which was also randomly assigned). Immediately following the auditory presentation of the stimuli, two boxes appeared on screen in random locations, each labelled with one of the plural suffixes (TEL or TOL). Participants were requested to click on the box corresponding to the plural form they had heard. If they provided an incorrect response, the trial was repeated to ensure that they had correctly heard the singular and plural forms of all exposure stimuli. This task was added during the exposure phase to ensure that participants were paying attention to the stimuli and had their speakers turned on. The random button position manipulation was chosen so that participants were required to actively seek out their response and could not repeatedly press the same button. We recorded the number of errors participants committed and used this information to exclude those who were either not paying attention or did not have proper audio equipment, as described in the Participants section below.

Immediately after this exposure phase, the first test phase began. Test trials consisted in the presentation of a stem followed by both possible plural forms (i.e., CVCV-tɛl and CVCV-tɔl) with an ISI of 500 ms. The order of presentation of the two plural forms was randomized across trials. After all three stimuli were played, two buttons appeared at random locations on the screen (exactly as during the exposure phase) labelled with the two plural suffixes. Participants were asked to select which of the two plural forms they thought was correct. They were first tested on the exposure set (i.e., the same items that they had just heard) and then on another randomly assigned set of 32 stimuli that they had not previously heard (e.g., they might have been exposed to and tested on set B and then tested on set A with set C being reserved for retest). They were not told anything about the two types of items they were tested on.

Approximately twelve hours after having completed the first session, participants received an email inviting them to log back into the website (the server refreshed every 10 minutes, so some participants received the email after 12 hours and 9 minutes, for example). They had exactly one hour to do so. If participants attempted to log in before the twelve hours had passed, they were instructed to return later. If they attempted to log in after the one-hour grace period, they were blocked from continuing and were excluded from the study. When they logged back in, they immediately began the retest. They were first tested on the exposure set again, and then on the third set of stimuli (i.e., the reserved set of items they had heard neither during exposure, nor during the first test). Participants were therefore tested on the exposure items and different sets of novel items at both test (session 1) and retest (session 2).

At the end of the retest, participants were asked to fill out a questionnaire concerning some basic personal information, as well as their sleep habits and strategies during the task.

On average, the first session (exposure + test) lasted around 20 minutes and the second session (test + questionnaire) 10 minutes.

2.1.3 Participants

Participants were American English speakers recruited on Amazon's Mechanical Turk platform. "Wake" batches were launched at 16:00 CET (10:00 EST) while "sleep" batches were launched at 04:00 CET (22:00 EST). This means that half of the participants were recruited in the morning, to return 12 hours later at the end of their day. The other half of participants were recruited in the evening, completing the first session before going to bed, and then participating in the second session the following morning.

Once recruited, participants were redirected from the Mechanical Turk website to our own website. Participants were randomly assigned to the harmony or disharmony condition upon logging into the website for the first time. After completing the first session, they received compensation for the "HIT", and if they completed the second session, they were awarded a bonus payment. A total of 723 participants were recruited, but 146 did not complete the first session, 297 did not return for the second session,⁶ and a further 10 completed either too few or too many trials (for instance, by doing the first session twice). Of the 245 participants who correctly completed both sessions, 72 were excluded from data analysis for the following reasons: made at least ten errors out of the 64 trials during the exposure phase (N=39), used only one response (i.e., either *tel* or *tol*) throughout an entire test phase (N=1), did not fill out questionnaire (N=16), napped during the day (N=11) or did not respond to question about napping (N=5), or took notes during exposure (N=16) or did not respond to the question about note taking (N=16).⁷ Note that some excluded participants fall into multiple categories (i.e., took notes during exposure *and* did not answer question about napping).

A total of 173 participants aged 20 to 67 (M=37; SD=10.0) were included in data analysis, distributed into the harmony/disharmony and wake/sleep conditions as shown in Table 1. Because of the difficulty in recruiting participants using Mechanical Turk who must return after a specific amount of time, we wound up launching a great many more 'sleep' batches, such that the total numbers of participants in the sleep groups were larger than those in the wake groups.

⁶ This was unexpected given that participants were told about the bonus and about the shorter duration of the second session from the outset. It might be because they had to start the second session between 12 and 13 hours after finishing the exposure phase (modulo the ten-minute refresh time of the server), and/or because the difference between the two payments was relatively small (US\$ 2.00 for the first session and US\$ 3.00 for the second one).

⁷ We decided to be conservative by excluding participants who may have taken notes or napped, erring on the side of caution when no information was provided. Note taking indeed undermines our ability to observe learning, and even daytime napping has been shown to consolidate motor-related memories (Lahl, Wispel, Willigens, & Pietrowsky, 2008; Nishida & Walker, 2007).

	Harmony	Disharmony
Wake	N = 26	N = 27
Sleep	N = 58	N = 62

Table 1: Distribution of participants into experimental conditions.

We compared the participants in the wake and sleep groups in a number of ways based on their responses to the questionnaire, as shown in Table 2. For most measures, the two groups did not differ. There was an expected asymmetry between the levels of fatigue at test and retest for the wake and sleep groups. The sleep group (initial test at the end of the day) reported an average higher level of fatigue at initial test than the wake group (initial test in the morning). Likewise, the wake group reported an average higher level of fatigue at their end of the day test (i.e., at retest) than the sleep group. Unsurprisingly, participants were thus more tired during the session they took part in at the end of the day.

	Wake	Sleep	Difference
Age	36.2	37.9	$t < 1$
Gender (ratio M:F)	0.89	1.11	$\chi^2 < 1$
Personality (ratio morning:evening)	0.77	0.50	$\chi^2 = 1.2, p > .1$
Concentration at test (1–5)	4.70	4.64	$t < 1$
Concentration at retest (1–5)	4.68	4.52	$t = 1.4, p > .1$
Fatigue before test (1–5)	1.55	1.80	$t = 2.0, p < 0.05$
Fatigue before retest (1–5)	2.22	1.58	$t = 3.5, p < .001$
Prior sleep quantity (hours)	6.91	6.79	$t < 1$
Prior sleep quality (1–5)	3.60	3.50	$t < 1$
Ideal amount of sleep (hours)	7.70	7.58	$t < 1$

Table 2: Average participant responses to questionnaire and test statistics comparing the wake and sleep groups.

2.2 Results

As a sanity check before beginning our planned analyses, we compared the number of ‘tel’ responses to the number of ‘tol’ responses. Recall that in the input, each suffix was used an equal amount of the time. On average, participants had no preference for one or the other ending (mean_{‘tel’} = 50.2%, SD = 8.6).

We calculated the accuracy of responses during the test phase for all participants. These data were analysed using logistic mixed-effects models in R (Bates, Maechler, Bolker, & Walker, 2014),

run with BOBYQA optimisation (Powell, 2009) whenever a model did not converge with maximal random effects structure. Of course, what was considered a ‘correct’ response depended on the rule participants were exposed to, such that harmonic and disharmonic responses were considered correct and incorrect, respectively, in the harmonic condition, and vice versa for the disharmonic condition. All factors in all of the models detailed below were defined using contrast coding, and significance was assessed through model comparison following the procedure described in Levy (2014), which entails the removal of factors and interactions always in comparison to the full model. Results for exposure items and novel items were analysed separately and are reported here sequentially. Good performance on the former can be reached either by learning the rule or by memorising the suffix for each individual item; good performance on the latter, by contrast, is an unambiguous indicator of rule learning.

We began by analysing the initial test session only, so as to compare our results with those reported in the literature. Indeed, data from the initial test session are qualitatively comparable to results from Pycha et al. (2003), in that they represent performance just after exposure. Mean accuracy from the initial test are displayed in Figure 1.

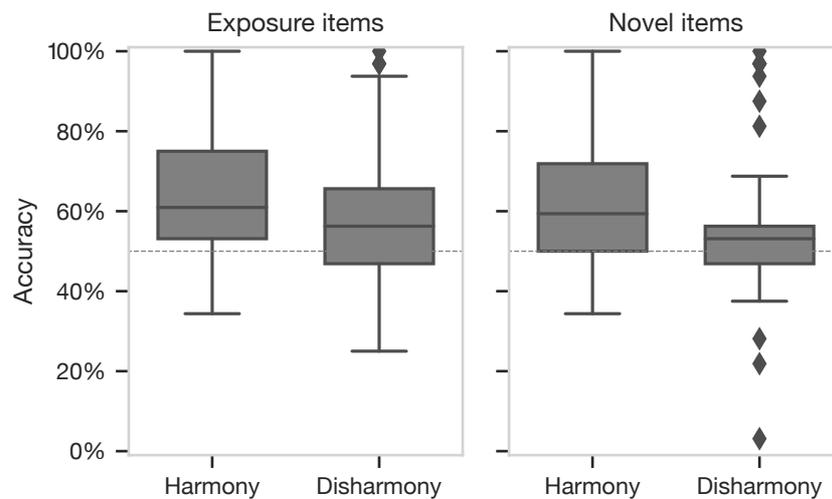


Figure 1: Boxplots showing mean accuracy scores at initial test on exposure and test items as a function of rule.

We designed models with fixed factors for Rule (harmony vs disharmony) and Interval (wake vs sleep), and their interaction, and random intercepts for Participant and Stem (one model for exposure items, and another for novel items). These models were compared to simpler models that excluded the factor Rule. Unlike Pycha et al. (2003), performance in the first test phase was better for harmony than for disharmony (exposure items: $\beta = 0.16$, $SE = 0.06$, $\chi^2(1) = 6.44$, $p < 0.02$; novel items: $\beta = 0.21$, $SE = 0.07$, $\chi^2(1) = 9.58$, $p < 0.002$). Thus, just after exposure, participants had learned the phonetically natural rule better than the unnatural one. However, no significant difference was

observed in the full model for the factor Interval or the interaction between Rule and Interval (both $z < 1$). In other words, whether participants were tested in the morning or in the evening did not influence the results in any way. The better performance on the harmony compared to the disharmony rule is also reflected by the percentage of participants who reached above-chance performance: According to a binomial test, the threshold for individual above-chance performance is 21 correct responses on the 32 exposure items (65.6%) and 39 on the 64 novel items (60.9%).⁸ For harmony, 42% and 37% of participants were at or above threshold on the exposure and novel items, respectively; for disharmony, this was the case for only 27% and 12% of participants, respectively.

We then analysed the full dataset, including both sessions; we first consider performance on the exposure items. Mean accuracy across the different manipulations for these items are shown in Figure 2.

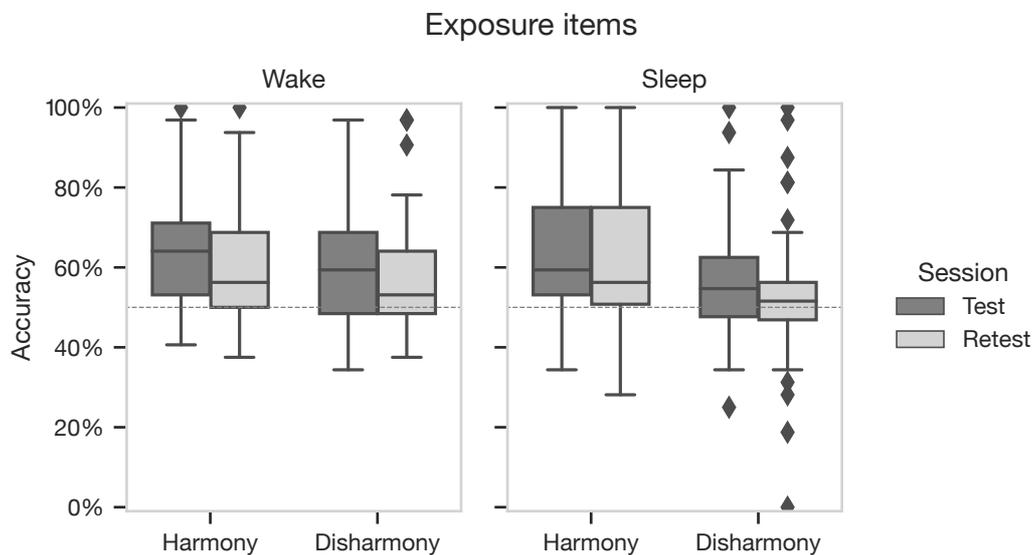


Figure 2: Boxplots showing mean accuracy on exposure items for the wake and sleep groups as a function of rule and test session.

A full model was designed that included the following fixed factors: Rule (harmony or disharmony), Interval (wake or sleep), Session (test or retest), and all the two- and three-way interactions. The random structure included intercepts for Participant and Stem, which both included random slopes for Session. This model was compared to simpler models that excluded one of the factors or interactions. The full model was found to explain significantly more variance than a model which excluded Rule ($\beta = 0.18$, $SE = 0.07$, $\chi^2(1) = 7.11$, $p < 0.008$), and one that excluded Session ($\beta = 0.06$, $SE = 0.03$, $\chi^2(1) = 4.15$, $p < 0.05$), but no such difference was observed for Interval or any of

⁸ Specifically, 21/32 yields a p-value of 0.055. Since 22/32 yields a p-value of 0.025 (well under 0.05) and that that threshold would force us to exclude even more participants from our secondary analysis, we have considered 21/32 correct responses to be representative of learning.

the interactions (all $z < 1$). Thus, performance was generally better for participants exposed to harmony than for those exposed to disharmony, and there was a general decrease in performance over the course of 12 hours, regardless of whether this period contained a night of sleep or not.⁹

We further examined to what extent participants' performance on individual exposure items was stable across the two test sessions. We thus designed a logistic mixed-effects model with performance on exposure items in Session 2 (retest) as the dependent variable, performance in Session 1 (test) as a fixed predictor, and a random intercept for Participant. Participants tended to respond correctly to the same exposure items at test and retest ($\beta = 0.86$, $SE = 0.06$, $\chi^2(1) = 198.57$, $p < 0.001$), indicating that they were likely relying on having memorised these items rather than extending the phonological rule to them as they would need to do for novel items.

We next considered the results for novel items. Mean accuracy across the different manipulations for novel items is shown in Figure 3.

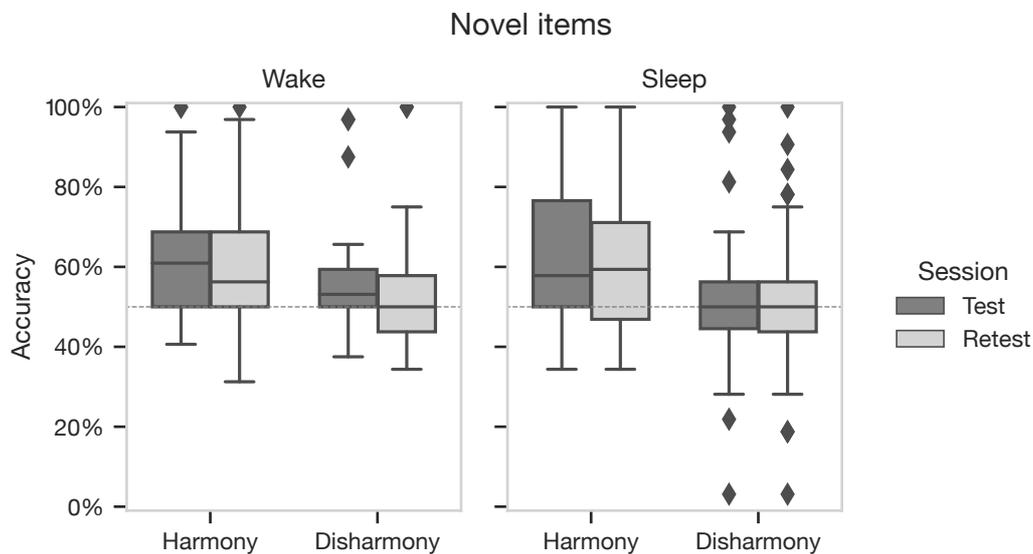


Figure 3: Boxplots showing mean accuracy on novel items for the wake and sleep groups as a function of rule and test session.

Models were designed identically to those for the exposure items, except that there were no slopes for session under Stem in the random structures, since the novel stems at test and retest were not the same (see section 2.1.2). The full model was found to explain significantly more variance than a model which excluded Rule ($\beta = 0.23$, $SE = 0.07$, $\chi^2(1) = 9.63$, $p < 0.002$), but not a model which excluded Session ($\beta = 0.03$, $SE = 0.02$, $\chi^2(1) = 1.30$, $p > 0.1$). For the factor Interval and all

⁹ We also examined if performance on exposure items was above chance level at both test and retest. At test, this was the case for all groups (wake/harmony: $z = 4.5$, $p < 0.0001$; wake/disharmony: $z = 3.0$, $p < 0.004$; sleep/harmony: $z = 5.8$, $p < 0.0001$; sleep/disharmony: $z = 3.7$, $p < 0.001$), while at retest it was the case for all groups except sleep/disharmony (wake/harmony: $z = 3.2$, $p < 0.002$; wake/disharmony: $z = 2.2$, $p < 0.03$; sleep/harmony: $z = 5.0$, $p < 0.0001$; sleep/disharmony: $z = 1.0$, $p > 0.1$).

interactions, no effects were observed in the full model (all $z < 1$).¹⁰ This indicates that harmony was generally learned better than disharmony, and that unlike for the exposure items, performance for neither rule decreased over the course of 12 hours, regardless of whether this period contained a night of sleep or not.

Overall performance was low, especially for the disharmony groups and even for the exposure items. The most obvious way to evaluate differences in learning over time, however, is to start from a base where all participants demonstrate individual learning and measure how their performance changes. We therefore ran a series of post-hoc analyses with the full dataset but including only those participants who achieved above-chance performance on exposure items in the initial test session (according to a binomial test; see above), thus demonstrating that they had at least learned how the rule applied to stems they had seen before. There were 59 such participants, aged 20 to 59 ($M = 39$; $SD = 10.0$). The distribution of these participants across experimental conditions is shown in Table 3.

	Harmony	Disharmony
Wake	N = 13/26	N = 9/27
Sleep	N = 22/58	N = 15/62

Table 3: Distribution of participants with above-chance performance on exposure items in the initial test into experimental conditions. Values are shown over the total number of participants included in the original analyses (cf Table 1).

Models were designed identically to those that we used to analyse the data from all participants. As before, we first consider performance on the exposure items. Mean accuracy across the different manipulations for these items are shown in Figure 4; note that as we only included participants who achieved above-chance performance in the initial test session, the dark grey boxes are all above chance level.

¹⁰ At test, performance was above chance on novel items for all groups except sleep/disharmony (wake/harmony: $z = 3.5$, $p < 0.001$; wake/disharmony: $z = 2.1$, $p < 0.04$; sleep/harmony: $z = 5.2$, $p < 0.0001$; sleep/disharmony: $z = 1.1$, $p > 0.1$). At retest, performance was above chance for the harmony but not the disharmony groups (wake/harmony: $z = 2.6$, $p < 0.009$; wake/disharmony: $z = 1.3$, $p > 0.1$; sleep/harmony: $z = 4.4$, $p < 0.0001$; sleep/disharmony: $z < 1$.)

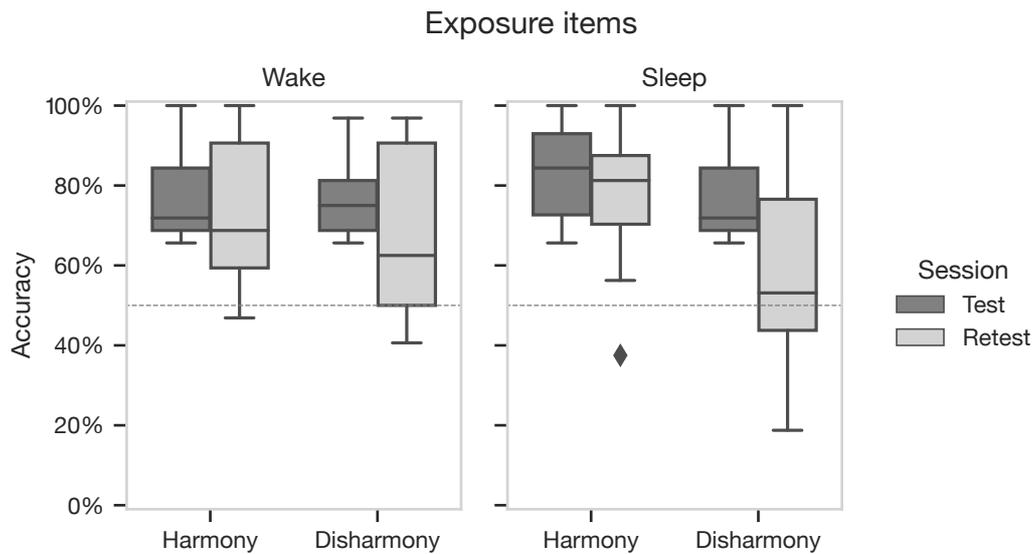


Figure 4: Boxplots showing mean accuracy on exposure items for the wake and sleep groups as a function of rule and test session, for participants with above-chance performance on exposure items at initial test.

The analyses revealed an effect of Session ($\beta = 0.15$, $SE = 0.06$, $\chi^2(1) = 4.77$, $p < 0.03$), with better performance in test than in retest, and an interaction between Rule and Session ($\beta = 0.12$, $SE = 0.06$, $\chi^2(1) = 4.28$, $p < 0.04$), but no other main effects or interactions (Rule: $\beta = 0.23$, $SE = 0.15$, $\chi^2(1) = 2.25$, $p > 0.1$; Rule \times Interval: $\beta = 0.19$, $SE = 0.15$, $\chi^2(1) = 1.65$, $p > 0.1$; all others: $z < 1$). The interaction between Rule and Session was due to the fact that participants exposed to harmony showed a smaller decrease in performance between test and retest than those exposed to disharmony.

Next, we consider performance on the novel items. Mean accuracy across the different manipulations for these items are shown in Figure 5.

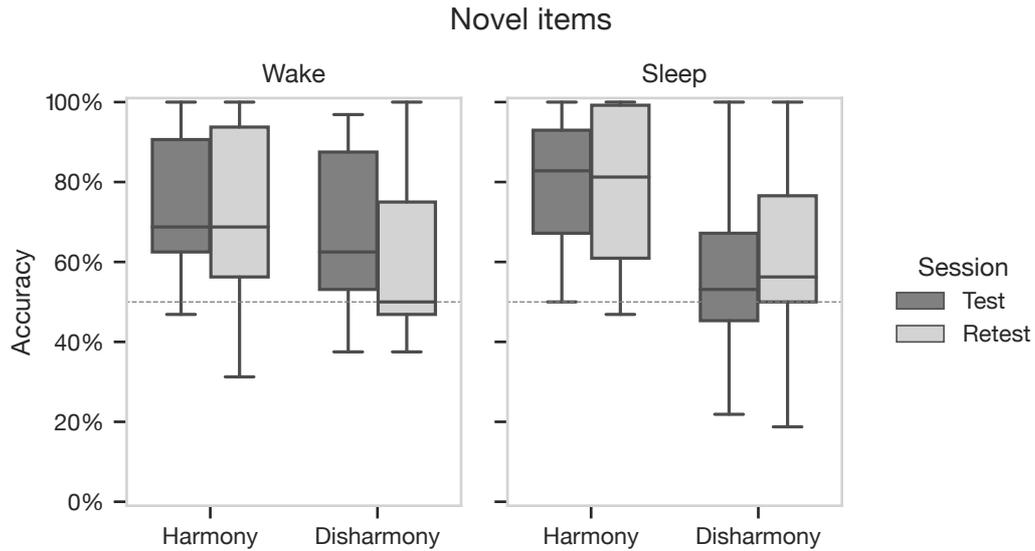


Figure 5: Boxplots showing mean accuracy on novel items for the wake and sleep groups as a function of rule and test session, for participants with above-chance performance on exposure items at initial test.

The analyses revealed a marginal effect of Rule (Rule: $\beta = 0.38$, $SE = 0.19$, $\chi^2(1) = 3.73$, $p = 0.053$), and no other main effects or interactions (Rule \times Interval: $\beta = 0.23$, $SE = 0.19$, $\chi^2(1) = 1.33$, $p > 0.1$; Session \times Interval: $\beta = 0.07$, $SE = 0.05$, $\chi^2(1) = 2.33$, $p > 0.1$; all others: $z < 1$).

Overall, the analyses on this restricted set of participants who reached above-chance performance on the exposure items in the initial test session confirm the bias in favour of harmony as opposed to disharmony. For the exposure items, they additionally show a larger decrease in performance between test and retest in the disharmony condition than in the harmony condition. Of course, these post-hoc analyses entailed a substantial reduction in sample size—to a third of the full sample—but they may serve to confirm what our full analysis already shows. That is, while we have no evidence of sleep-related consolidation in our experiment, a consistent bias favouring harmony over disharmony is present in all of our analyses.

3 Discussion

This study considered the hypothesis that phonetically natural rules might benefit from a bias during learning such that they are more likely to survive the repeated transmission process than phonetically unnatural rules. Previous examinations of this hypothesis have indeed shown easier or better learning of phonetically natural rules compared to unnatural ones, though the natural rules are almost always less formally complex than the unnatural ones (for a review, see Moreton & Pater, 2012). Concerning the few studies that tested rules that are logically equivalent in terms of complexity, methodological concerns (e.g., predisposition to the natural rule because of L1

experience, low statistical power) mean that evidence of a learning bias favouring phonetically natural rules is yet weak. The present study focusses on two phonological rules, one natural and typologically recurrent (vowel harmony), one unnatural and exceedingly rare (vowel disharmony), matched in complexity. As weak learning biases are not easy to demonstrate in the lab and given that at least two previous studies who also considered these rules did not show a difference in learning patterns (Pycha et al., 2003; Skoruppa & Peperkamp, 2011), we further examined the role of sleep-related memory consolidation. Indeed, we did not necessarily expect to observe a clear asymmetry between harmony and disharmony at initial test. Rather, we expected that we might observe a numerical trend (like Pycha et al.), and that this small bias might be reinforced by differential memory consolidation after sleep, yielding a boost for the phonetically natural but not the unnatural rule. Our findings do not entirely align with these predictions, but rather highlight the existence of a phonetically-motivated learning bias. Contrary to previous work, we found that the phonetically natural and typologically common rule of vowel harmony was learned better than the phonetically unnatural and exceedingly rare rule of vowel disharmony. However, the extent to which the rule was learned and applied to novel items remained stable over a period of 12 hours, regardless of the absence versus presence of a night of sleep. We will discuss these findings in turn.

Better learning of vowel harmony in the first test session, hence regardless of sleep, contrasts with the two previous studies just mentioned (Pycha et al., 2003; Skoruppa & Peperkamp, 2011). Here, we observed better performance for the harmony than for the disharmony rule during the initial test session for both exposure and novel items. We further observed a clear effect of harmony versus disharmony when considering the full data set (i.e., including the retest session), again both for exposure and novel items. Participants exposed to the phonetically natural harmony rule performed better than those exposed to the unnatural disharmony rule. This, then, is the first solid evidence that harmony is easier to learn than disharmony.

Our sample was considerably larger than those in previous studies, which might have been underpowered to detect a learning bias. Another methodological difference worth mentioning is that our participants were native speakers of *English* tested on *French* stimuli. Previous research has already shown that artificial language learning experiments focussing on phonology can be run with non-native stimuli (Steele et al., 2015). It could be, though, that such stimuli are actually *more* appropriate for exploring phonological learning asymmetries, given the hypothesis that their processing is both more phonetic and less likely to be influenced by orthographic knowledge. We leave this to future work, but note the use of non-native stimuli as a potentially crucial manipulation in the present article, allowing us to uncover a learning bias.

Yet, before uncorking the champagne and toasting to phonetic substance in phonological grammar, it is important to consider a plausible alternative explanation for the bias favouring harmony over disharmony. Although English does not have a vowel harmony rule, it is possible that the English lexicon happens to be organised in such a way that words are more likely to be harmonic

than disharmonic. If that is true, participants in our experiment may bring their knowledge of English (which biases them towards harmony) and extend that knowledge to the artificial language. At first sight, one way to test for this bias would be by conducting a control experiment without an exposure phase. Thus, participants would be tested on their relative preference for harmony versus disharmony rather than on their relative capacity to learn the rules. Note, though, that finding a preference for harmony over disharmony prior to exposure to the pattern in an artificial language would have no consequences for our findings or conclusions. Indeed, general phonetic knowledge (of the kind we hypothesise to be present in phonological grammar) could lead participants to prefer harmonic words over disharmonic ones, and be precisely the driving force that renders the learning of a harmonic pattern easier than the learning of a disharmonic pattern. In other words, a simple preference experiment would not allow us to understand whether the learning bias we found stems from participants' experience with their native language. Instead, to assess this specific possibility, we conducted a series of analyses aimed at calculating the amount of evidence for harmony compared to disharmony present in the English lexicon. This allows us to measure how much the bias found in our experiment could plausibly be due to knowledge of English. We considered two possible sources of a lexical harmony bias: 1) a high relative amount of harmonic words (i.e., words containing only front or only back vowels), and 2) a relative amount of harmonic words that is larger than would be expected by chance, given the frequency of the vowels of English.

We extracted all polysyllabic lemma forms from the CMUdict (*CMU Pronouncing Dictionary*, 2008). (Monosyllabic words are not informative with regards to vowel harmony, and all such words were therefore excluded from the present analysis.) For each word, a harmony score was calculated, where each vowel within the word was assigned a score, 0 or 1, depending on whether it was a front or back vowel, respectively. Schwas were not considered in the analysis, as it is unclear whether they should be considered as front or back. A word's harmony score was considered as the variance of backness of its vowels, such that a word with three front vowels (0, 0, 0) or three back vowels (1, 1, 1), would have a harmony score of 0 (no variance in backness), while a word with one front and two back vowels (0, 1, 1) would have a harmony score of 0.22. This gives a distribution of harmony scores bounded between 0 and 0.25 (regardless of word length), with a word containing as many front as back vowels) having a score of 0.25. The distribution for words in English can be seen in the top left panel of Figure 6.

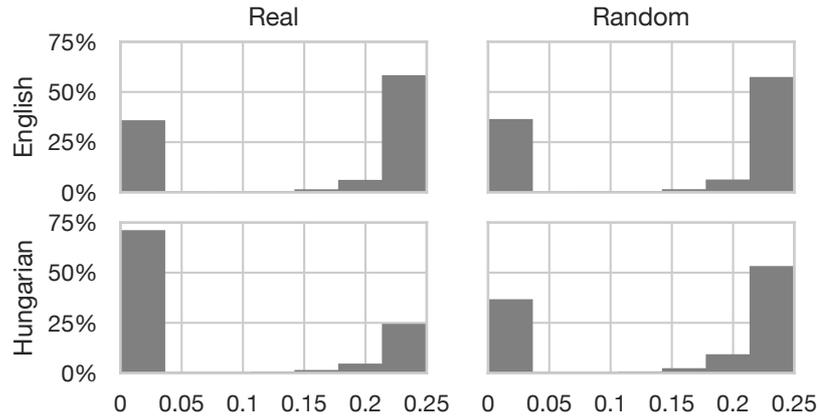


Figure 6: Distributions of palatal harmony scores for real and random lexicons in English (no vowel harmony rule) and Hungarian (productive vowel harmony rule). Each bar represents the percentage of the lexicon that falls into that bin. A score of 0 means a word is fully harmonic and a score of 0.25 means a word contains as many front as back vowels.

Figure 6 shows that only around 35% of English polysyllabic words are harmonic.¹¹ To ensure that this is not driven by a systematic pressure that participants may be able to pick up on, we generated a random version of English by extracting all vowels from all words, shuffling them, and reinserting them into the consonant frames. A word like *moodiness* - /mudines/ could therefore become /mædunəs/ or /mɪdʌnis/, etc. We then recalculated the distribution of harmony scores for this random English lexicon and compared it to the real English lexicon. As can be seen in the top right panel of Figure 6, the random distribution is almost completely identical to the real distribution; in fact, only 1.74% of the real distribution is not overlapping with the random one.

To ensure that our measure properly captures explicit harmony, we also performed the same analysis on the lexicon of a language with productive vowel harmony: Hungarian. The real Hungarian lexicon should contain more harmonic words than a random lexicon made by shuffling the vowels, because Hungarian words typically contain only front or back vowels, a grammaticalised pressure in the language which, while known to target morphophonological alternations like in example (1) from the Introduction, also affects roots (note that both roots in (1) are themselves harmonic). We used the Hungarian pronunciation dictionary (Grimes, 2006) to extract the phonological transcription of words. We followed the traditional divide of Hungarian vowels, with /y/, /y:/, /ɛ/, /e:/, /ø/, and /ø:/ considered front, /u/, /u:/, /o/, /o:/, /ɒ/, and /a:/ considered back, and /i/ and /i:/ considered neutral. As can be seen

¹¹ The fact that words appear to either be fully harmonic or maximally disharmonic (i.e., either have a harmony score of 0 or 0.25) is due to the high frequency of shorter words (meaning our sample contains a great many short words and relatively fewer long words). Indeed, the median number of vowels in the polysyllabic English words is 2.0; thus, if those two vowels have different values for backness, the word is maximally disharmonic. It is not the case that very many long English words are maximally disharmonic. For instance, the average harmony score for English words with more than four syllables is 0.206.

in the bottom left panel of Figure 6, the Hungarian lexicon is overwhelmingly harmonic, with over 70% of words containing only front or back vowels. We then tested that this distribution is a result of an explicit pressure by creating a random version of the Hungarian lexicon, just as we did for English above. As can be seen in the bottom right panel of Figure 6, a random Hungarian lexicon shows a great many more disharmonic words, and far fewer harmonic words; the bar on the left is nearly halved, with only 36% of the random words containing only front or only back vowels, compared to 70% in the real lexicon. Our method therefore appears able to distinguish between harmony by chance (due to the statistical distribution of vowels) and systematic harmony (due to an active process in the language).¹²

Our reasoning has considered two ideas: the presence of fully harmonic words (made up of only front or only back vowels), and a pressure for the vowels to cluster in a certain way above what would be expected by chance, given their individual frequencies. While it may be enough to consider that the number of harmonic words in English does not provide speakers with enough evidence of harmony to explain our experimental results (only 35% of English words are harmonic), our lexical analyses further considered the possibility that despite the relatively few harmonic words in English, there are still more harmonic words than would be predicted by chance. This also does not seem to be the case in the English lexicon (though it very clearly is in the Hungarian lexicon). All in all, we conclude that the experimental results reported above cannot simply be explained by a language-specific bias, whereby learners fare better on whatever pattern is most frequent in their native lexicon.

Our results strongly suggest, then, that vowel harmony has a preferential status in learning compared to vowel disharmony. We argue that this is due to the fact that vowel harmony is phonetically natural (grounded in phonetic substance) where vowel disharmony is not. This type of phonetically-based bias could play a role in explaining why phonetically unnatural patterns like vowel disharmony are rare in the world's languages. Such a bias could come into play, for example, in the case of ambiguous input, with multiple plausible representations. In such a case, a learner could use their phonetic knowledge to determine that one representation is more likely than another. This type of substantive bias could instantiate itself by complementing what Blevins (2004) refers to as CHANCE, where one representation is randomly chosen over another. In cases where competing representations are equally likely (i.e., phonetically plausible), a learner might indeed assume one at the same rate as another. However, in cases where one representation is phonetically more natural than a competing one, a phonetically-based learning bias, encoded in synchronic grammar, might lead a learner to favour the former. This would explain, in part, why such phonetically-grounded (natural) patterns are typologically more frequent than their unnatural counterparts. While diachronic

¹² The Hungarian pronunciation dictionary does not provide morphological information, which means that we also analysed a certain number of compound words. As vowel harmony does not spread across the morpheme boundary within free morpheme compounds, our analysis underestimates the extent to which the Hungarian lexicon is harmonic.

explanations for sound patterns have a great deal of explanatory power, dismissing synchronic explanations entirely is overly restrictive, especially in light of our experimental results.

A further point to think about, given previous work highlighting the importance of simplicity bias in language (i.a., Culbertson & Kirby, 2016; Moreton & Pater, 2012a), is how we have considered complexity. The rules we tested were of equal complexity (i.e., they were both defined by a single feature), but could asymmetries amongst specific phonological features also affect the perceived complexity (and learnability) of patterns generally? A number of previous studies have shown that a difference in one featural dimension is not necessarily perceived as equal to a difference in another featural dimension, and that this can be the result of both universal phonetic properties and language-specific ones (Cole, Jakimik, & Cooper, 1978; Ernestus & Mak, 2004; Martin & Peperkamp, 2017; White, 2014). In our design this is not of great import, since both our harmony and disharmony patterns involve the same features; only the direction of the alternation changes. But it is interesting to consider if a phonological pattern involving a perceptually distinct feature would be more difficult to learn than one involving a feature with less perceptually distinctive phonetic correlates. This is precisely the argumentation provided by White (2014) to explain why in his study participants more readily learned a voicing alternation than a manner of articulation alternation (though both involved only one feature). If learners use their phonetic knowledge in learning phonological alternations, such differences are to be expected, and could even depend on modality. That is, an alternation that is more distinct articulatorily might be harder to learn if the task requires the participants to themselves produce the items.

Turning finally to the learning we both did and did not observe, the design we used allows us to highlight the fact that performance on exposure items and that on novel items likely involve different underlying mechanisms: memorisation on the one hand, and rule learning on the other. Indeed, performance on individual exposure items correlated between the first and the second test session; that is, the items that participants performed well on during retest tended to be ones they had also performed well on during the initial test. Moreover, even though we did not perform any statistical analyses to compare performance on exposure and novel items, it is worth noting that for the former, performance generally worsened between test and retest whereas for the latter it neither improved nor deteriorated.

The finding that harmony is learned better than disharmony made it more difficult to study our second question (i.e., whether the absence vs presence of sleep would differentially affect learning), since baseline performance before the 12-hour interval was different for participants exposed to harmony and those exposed to disharmony. The number of participants overall who reached above-chance performance during the first session was, moreover, low. Yet, our post-hoc analyses on the subset of participants with above-chance performance on exposure items in the initial session showed the same results as the analyses on the entire sample: Performance on novel items was stable across the two test sessions, and this was true for both the harmony and disharmony conditions.

Thus, we observed no evidence of consolidation at all. If consolidation aided generally, without differentially benefitting one of the two rules we tested, we should have seen an interaction between Session and Interval, but we observed no such interaction. We also did not observe a triple interaction, where sleep benefits one of the rules more than the other. However, the fact that performance on novel items was stable across the two test sessions is a robust result, and the first one to show that phonological rule learning in an artificial language learning paradigm is not entirely ephemeral. Of course, it would be interesting to see how long exactly the learning effect survives.

Although we found no evidence for consolidation, it would be premature to claim that phonological rule learning cannot benefit from sleep-related memory consolidation, especially given evidence for such consolidation in other cases of linguistic learning, such as the learning of novel phonotactic constraints (Gaskell et al., 2014) or morphosyntactic rules (Batterink et al., 2014).¹³ In light of our use of non-native stimuli, it is also worth noting that in the domain of word learning, sleep-related consolidation has been shown to play a special role for words containing non-native sounds (Havas et al., 2018). The absence of consolidation in our study is, as any null effect, particularly hard to interpret. There are at least two aspects of our manipulation that might have impacted our ability to observe consolidation. First, the population we tested was on average middle-aged (mean age = 37), and memory consolidation tends to decline after young adulthood (Scullin & Bliwise, 2015). Second, performance was generally fairly low (even though above chance level in the first session for all groups). It is worth noting that both Gaskell et al. (2014) and Batterink et al. (2014) tested young adults with tasks on which the participants reached high accuracy. It would thus be pertinent for future research to lengthen the exposure phase and/or use a phonological rule that is more easily learnable (while still being difficult enough to observe variation in the population), as well as to select participants in their early twenties.

If in future experiments we do see a difference in performance according to the presence vs absence of sleep in between testing sessions, the causal role of sleep would still need to be confirmed. Indeed, rather than the presence of sleep per se, it might be the absence (or the reduced amount) of language processing during sleep that accounts for such difference. A reliable way to demonstrate the role of sleep consists of recording EEG (electroencephalography) during sleep and carrying out individual correlation analyses; that is, sleep-dependent consolidation is revealed by a positive correlation between the amount of REM sleep and/or slow-wave sleep (SWS) on the one hand and the improvement in performance following a period of sleep on the other hand (Batterink et al., 2014; Gaskell et al., 2014; Nishida & Walker, 2007; Tamminen, Payne, Stickgold, Wamsley, & Gaskell, 2010; Walker & Stickgold, 2004). Thus, while for practical reasons we opted for online testing, it would eventually be necessary to resort to a lab-based study.

¹³ Note though, that Gaskell et al. (2014) used a production task, which involves procedural memory; consolidation for declarative memory such as the one involved in Batterink et al. (2014) and the present study has been argued to rely on different neural mechanisms (Diekelmann, Wilhelm, & Born, 2009).

A final methodological note warrants stating. While we had hoped that testing participants online would ease the difficulties of bringing participants to the lab at a precise 12-hour interval, we found that our strict test/retest setup was not ideal for Mechanical Turk. In the end, we wound up needing to recruit an excessively large number of participants simply because many did not return within the specified window. This was despite explicit instruction, monetary incentive (the second session paid a bonus that was larger than the original payment), and precisely timed automated reminder emails. We thus caution use of this platform in cases where participation in multiple test sessions depends on precise timing.

4 Conclusion

To conclude, we investigated whether phonetically natural rules benefit from a learning bias, and, if so, whether this bias stems from, or is enhanced by, consolidation during sleep. The results provide clear evidence in favour of a learning bias, but no evidence for a role of sleep. In fact, sleep did not appear to enhance learning at all, regardless of whether the rule to be learned was phonetically natural or unnatural. The question of whether or not sleep-dependent consolidation plays a role in shaping sound patterns in human language thus remains open. However, our results clearly demonstrate that a phonetically natural rule is learnt better than a phonetically unnatural rule. For the first time, we have clear evidence that vowel harmony (phonetically natural) is easier to learn than vowel disharmony (phonetically unnatural). This in turn may well affect the way such systems evolve during the repeated transmission process, influencing the asymmetrical distribution of phonetically natural rules compared to unnatural ones in the typology.

Acknowledgements

This work was supported by ANR-17-CE28-0007-01 and ANR-17-EURE-0017. We would like to thank Michel Dutat for setting up the server from which the experiment was run, and Page Piccinini for launching a number of the sleep batches.

References

- Archangeli, D., & Pulleyblank, D. (1994). *Grounded Phonology*. Cambridge, Massachusetts: MIT Press.
- Baer-Henney, D., Kügler, F., & van de Vijver, R. (2014). The interaction of language-specific and universal factors during the acquisition of morphophonemic alternations with exceptions. *Cognitive Science*, 1–33. <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.
- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*.
- Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, 65(2014), 169–179. <https://doi.org/10.1016/j.neuropsychologia.2014.10.024>
- Beguš, G. (2018). *Bootstrapping sound changes*. Retrieved from <https://ling.auf.net/lingbuzz/004299>
- Blevins, J. (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. <https://doi.org/10.1017/CBO9780511486357>
- Carpenter, A. C. (2010). A naturalness bias in learning stress. *Phonology*, 27(03), 345–392. <https://doi.org/10.1017/S0952675710000199>
- Carpenter, A. C. (2016). The Role of a Domain-specific Language Mechanism in Learning Natural and Unnatural Stress. *Open Linguistics*, 2(1). <https://doi.org/10.1515/opli-2016-0006>
- CMU Pronouncing Dictionary. (2008). Carnegie Mellon University.
- Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, 64(1), 44–56. <https://doi.org/10.1121/1.381955>
- Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16), 5842–5847. <https://doi.org/10.1073/pnas.1320525111>
- Culbertson, J., & Kirby, S. (2016). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01964>
- Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, 21(4), 803–820. <https://doi.org/10.1162/jocn.2009.21059>
- Dieckmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, 13(5), 309–321. <https://doi.org/10.1016/j.smrv.2008.08.002>
- Docherty, G. (1992). *The Timing of Voicing in British English Obstruents*. Berlin: Foris.
- Donegan, P. J., & Stampe, D. (1979). The Study of Natural Phonology. In D. A. Dinnsen (Ed.), *Current Approaches to Phonological Theory* (pp. 126–173). Bloomington & London: Indiana University Press.
- Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, 18(1), 35–39. <https://doi.org/10.1111/j.1467-9280.2007.01845.x>
- Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, 41(6), 1680–1695. <https://doi.org/10.1037/xhp0000113>
- Ernestus, M., & Mak, W. M. (2004). Distinctive phonological features differ in relevance for both spoken and written word recognition. *Brain and Language*, 90(1–3), 378–392. [https://doi.org/10.1016/S0093-934X\(03\)00449-8](https://doi.org/10.1016/S0093-934X(03)00449-8)
- Eszkénazi, M., Levow, G.-A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.
- Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, 425(October), 614–616. <https://doi.org/10.1038/nature01951>
- Finley, S. (2012). Typological asymmetries in round vowel harmony: Support from artificial grammar learning. *Language and Cognitive Processes*, 27(10), 1550–1562. <https://doi.org/10.1080/01690965.2012.660168>
- Finley, S., & Badecker, W. (2008). Analytic biases for vowel harmony languages. In N. Abner & J. Bishop (Eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics* (pp. 168–176). Somerville, MA: Cascadilla Proceedings Project.
- Finley, S., & Badecker, W. (2009a). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3), 423–437. <https://doi.org/10.1016/j.jml.2009.05.002>
- Finley, S., & Badecker, W. (2009b). Right-to-left biases for vowel harmony: Evidence from artificial grammar. *Proceedings of the 38th North East Linguistic Society Annual Meeting*, 1, 269–282.
- Gaskell, M. G., Warker, J. A., Lindsay, S., Frost, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep Underpins the Plasticity of Language Production. *Psychological Science*, 25(7), 1457–1465. <https://doi.org/10.1177/0956797614535937>
- Grimes, S. M. (2006). On the creation of a pronunciation dictionary for Hungarian. *Proceedings of the 6th Midwest Computational Linguistics Colloquium*. Presented at the Indiana University. Indiana University.
- Havas, V., Taylor, J., Vaquero, L., de Diego-Balaguer, R., Rodríguez-Fornells, A., & Davis, M. H. (2018).

- Semantic and phonological schema influence spoken word learning and overnight consolidation. *Quarterly Journal of Experimental Psychology*, 71(6), 1469–1481. <https://doi.org/10.1080/17470218.2017.1329325>
- Hayes, B. P., & Steriade, D. (2004). The phonetic basis of phonological Markedness. In B. P. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically Based Phonology* (pp. 1–33). Cambridge: Cambridge University Press.
- Hochmann, J.-R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, 177, 49–57. <https://doi.org/10.1016/j.cognition.2018.04.005>
- Hooper, J. B. (1976). *An Introduction to Natural Generative Phonology*. New York: Academic Press.
- Kimper, W. (2016). Asymmetric Generalisation of Harmony Triggers. In G. Ó. Hansson, A. Farris-Trimble, K. McMullin, & D. Pulleyblank (Eds.), *Proceedings of the 2015 Annual Meeting on Phonology*.
- Krämer, M. (1999). *A Correspondence Approach to Vowel Harmony and Disharmony*.
- Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, 17(1), 3–10. <https://doi.org/10.1111/j.1365-2869.2008.00622.x>
- Levy, R. (2014). *Using R formulae to test for main effects in the presence of higher-order interactions*.
- Martin, A., Abels, K., Adger, D., & Culbertson, J. (2019). Do learners' word order preferences reflect hierarchical language structure? In A. K. Goel, C. M. Seifart, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (pp. 2303–2309). Montreal, Canada: Cognitive Science Society.
- Martin, A., & Peperkamp, S. (2017). Assessing the distinctiveness of phonological features in word recognition: Prelexical and lexical influences. *Journal of Phonetics*, 62, 1–11. <https://doi.org/10.1016/j.wocn.2017.01.007>
- Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, 25(1), 83–127. <https://doi.org/10.1017/S0952675708001413>
- Moreton, E., & Pater, J. (2012a). Structure and substance in artificial-phonology learning, part I: Structure. *Language and Linguistics Compass*, 6(11), 686–701. <https://doi.org/10.1002/lnc3.363>
- Moreton, E., & Pater, J. (2012b). Structure and Substance in Artificial-Phonology Learning, Part II: Substance. *Language and Linguistics Compass*, 6(11), 702–718. <https://doi.org/10.1002/lnc3.366>
- Myers, S., & Padgett, J. (2014). Domain generalization in artificial language learning. *Phonology*, 31(3), 399–433. <https://doi.org/10.1017/S0952675714000207>
- Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. *PLoS ONE*, 2(4). <https://doi.org/10.1371/journal.pone.0000341>
- Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical Linguistics: Problems and Perspectives* (pp. 235–278). London: Longman.
- Ohala, J. J. (1994). Towards a universal, phonetically-based, theory of vowel harmony. *Proceedings of the 3rd International Conference on Spoken Language Processing*, 491–494. Yokohama, Japan.
- Peperkamp, S., Skoruppa, K., & Dupoux, E. (2006). The Role of Phonetic Naturalness in Phonological Rule Acquisition. *Proceedings of the 30th Annual Boston University Conference on Language Development*. Cascadia Press.
- Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (p. 39). Department of Applied Mathematics and Theoretical Physics, Cambridge University.
- Pycha, A., Nowak, P., Shin, E., & Shosted, R. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In M. Tsujimura & G. Garding (Eds.), *WCCFL 22 Proceedings* (Vol. 22, pp. 101–114). Somerville, MA.
- Schane, S. a., Tranel, B., & Lane, H. (1974). On the psychological reality of a natural rule of syllable structure. *Cognition*, 3(4), 351–358. [https://doi.org/10.1016/0010-0277\(74\)90004-3](https://doi.org/10.1016/0010-0277(74)90004-3)
- Scullin, M. K., & Bliwise, D. L. (2015). Sleep, Cognition, and Normal Aging: Integrating a Half Century of Multidisciplinary Research. *Perspectives on Psychological Science*, 10(1), 97–137. <https://doi.org/10.1177/1745691614556680>
- Skoruppa, K., Lambrechts, A., & Peperkamp, S. (2011). The role of phonetic distance in the acquisition of phonological alternations. *Proceedings of the 39th Annual Meeting of the North East Linguistic Society*, 2, 717–729.
- Skoruppa, K., & Peperkamp, S. (2011). Adaptation to novel accents: Feature-based learning of context-sensitive phonological regularities. *Cognitive Science*, 35(2), 348–366. <https://doi.org/10.1111/j.1551-6709.2010.01152.x>
- Steele, A., Denby, T., Chan, C., & Goldrick, M. (2015). Learning Non-Native Phonotactic Constraints Over the Web. *Proceedings of the 18th International Conference of the Phonetic Sciences*, 1–5.
- Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *The Journal of Neuroscience*,

- 30(43), 14356–14360. <https://doi.org/10.1523/JNEUROSCI.3028-10.2010>
- Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron*, 44(1), 121–133. <https://doi.org/10.1016/j.neuron.2004.08.031>
- White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition*, 130(1), 96–115. <https://doi.org/10.1016/j.cognition.2013.09.008>
- Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5), 945–982.