Salient dimensions in implicit phonotactic learning

Elise Michon[1,2,3]*, Emmanuel Dupoux*[3]*, Alejandrina Cristia*[3]

[1]Boğaziçi University, Turkey, [2]Ecole Polytechnique, France
[3]Laboratoire de Sciences Cognitives et Psycholinguistique, (ENS, EHESS, CNRS)
Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, France
michon.elise@gmail.com, emmanuel.dupoux@gmail.com, alecristia@gmail.com

Abstract

Adults are able to learn sound co-occurrences without conscious knowledge after brief exposures. But which dimensions of sounds are most salient in this process? Using an artificial phonology paradigm, we explored potential learnability differences involving consonant-, speaker-, and tone-vowel co-occurrences. Results revealed that participants, whose native language was not tonal, implicitly encoded consonant-vowel patterns with a high level of accuracy; were above chance for tone-vowel co-occurrences; and were at chance for speaker-vowel co-occurrences. This pattern of results is exactly what would be expected if both language-specific experience and innate biases to encode potentially contrastive linguistic dimensions affect the salience of different dimensions during implicit learning of sound patterns.

Index Terms: phonotactics, artificial grammar learning, talker identity, tones, consonants, vowels, internet-based testing

1. Introduction

All languages have sound patterns and, to acquire their native language, infants gradually learn to identify, combine and exchange its sound patterns. Adults keep some of these phonological abilities in order to speak their language(s), stay flexible to its evolution (appearance of loanwoards or new words, adaptation to different accents) and learn additional languages. But what are the dimensions along which sound patterns are computed? That is, how do learners encode co-occurrences that are frequent in their input?

In order to isolate the cognitive biases involved in the extraction and encoding of phonotactic patterns, artificial grammars can be built to contain specific regularities, for example /p/ always followed by /a/ and /d/ by /e/. Infants or adults are then exposed to these sound patterns in the laboratory, and subsequently tested on their implicit learning of the regularities. We consider that a regularity is learnt if participants react differently to items that are "legal", that is to say corresponding to the constraint they had been exposed to, versus "illegal", violating the frequent regularity – regardless of whether the item itself has been heard previously or not.

Although adults are able to learn new constraints on sound patterns in artificial grammars presented in the laboratory [1], not all constraints are equally easy to learn. One recurrent question in this line of research concerns which dimensions are readily encoded and which are less salient during phonological learning. For example, Onishi and colleagues [2] documented that a frequent association between a set of consonants and a vowel was reflected in the listeners' latency to repeat syllables, while an association between a set of consonants and the speaker's gender (presented for the same length of time and in the same manner) was not.

In other words, consonant-vowel co-occurrences were more salient than speaker-consonant co-occurrences.(We note briefly that this specific set of results could have been attributed to the task: the participants repeated the consonant-vowel co-occurrences in their own production, but they could not mimic the speaker-consonant co-occurrences. Nonetheless, as we will see below, Onishi and colleagues' conclusions actually hold true for a purely perceptual task.)

One possible explanation for this pattern of results is that participants ignore talker identity because they have learned, through experience, that this is not a linguistically relevant dimension; a second option is that this reflects an experience-independent bias. Indeed, vowels and consonants are involved in lexical contrasts in all languages and are therefore linguistically relevant units. In contrast, the speaker's voice is thought to be only an indexical cue (a pointer to the identity of the speaker), and is never linguistically relevant (i.e., lexically contrastive) in any human language.

To decide whether experience plays a key role, we turn to a third dimension. Variation of pitch in association with a syllable, called "tone" in the rest of the article, can, in fact, serve both of these roles: in tonal languages, like Mandarin Chinese, tones can be used to contrast words on their meaning. In non-tonal languages, such as English, these variations are better designated as intonation and they convey the emotions or the attitude of the speaker, for instance a pitch rise generally indicates a question while a pitch fall indicates a statement.

We built stimuli to test three types of associations:

(a) consonant-vowel: both are lexically contrastive in all languages

(b) speaker-vowel: the former is not lexically contrastive in any language

(c) tone-vowel: the former is lexically contrastive in some languages (so-called tone languages)

Our hypothesis is that associations between lexically contrastive dimensions should be easier to encode than associations between those that are lexically contrastive and those that are not. If linguistic relevance is determined by lexical experience, then consonant-vowel and tone-vowel co-occurrences will be equally learnable only for tone speakers; if not, then even speakers of a non-tonal language will more readily encode both consonant-vowel and tone-vowel co-occurrences than speaker-vowel co-occurrences. While we intended to compare speakers of tonal and non-tonal languages, we only had data from 16 tonal participants (divided into 3 conditions). We concentrate here on speakers of non-tonal languages, who nonetheless allow

us to assess whether language experience with tone is a necessary condition to render this dimension salient during phonological learning.

2. Methods

The present study was pre-registered on https://osf.io/kcf6w/. We followed the pre-registered procedure to the letter, other than being unable to investigate learning among participants with a tonal native language. The data from non-tonal participants is available from the project's osf site.

The experimental platform, hosted on the Ibex farm (http://spellout.net/ibexfarm/), has been programmed with language and tools proposed by Ibex, presenting items in a random order within each phase. Times of exposure for sounds and images resulted from previous knowledge in experimental phonology added to cross-browser testing. Instructions have been translated in three languages: French, English, Mandarin Chinese and revised by native speakers. Random selection of the condition, rule and order of the lists, and redirection to the corresponding experiment setup was performed by a PHP script. A final form recorded useful information about the participant to distinguish groups in the statistical analysis.

2.1. Stimuli

All stimuli can be represented as follows: $C_1VC_2 - T - S$, meaning a consonant-vowel-consonant, spoken with a given Tone by a given Speaker. Stimuli were recorded from Vietnamese speakers, in Paris, France, where a vibrant community speaking this tonal language can be found. Further details follow on each of the elements of the stimuli.

- The first consonant C_1 could be either /f/ or /s/: both are voiceless fricatives but differ in place of articulation (labio-dental vs. alveolar). For the consonant-vowel co-occurrence condition, each consonant could only be combined with half of the vowels; for both of the other conditions, they combined freely.
- The vowel V was always restricted by something in the context (by the consonant, by the tone, or by the speaker), such that, depending on the context, it had to be chosen from one of 2 phonological classes:
  - front unrounded vowels, with variations in tongue height: close /i/, close-mid /e/ and open-mid /ɛ/.
  - back rounded vowels, with variations in tongue height: close /u/, close-mid /o/, open-mid /ɔ/.
- The last consonant C_2 closed the syllable and introduced variability in the items heard by the participant without having an influence on other factors of the syllable. That is, each coda co-occurred with every other CVC-T-S combination. Following the restrictions in Vietnamese, it could be any of /p/, /t/, /k/, /m/, /n/, and /ŋ/.
- The tone T could be one of the 2 allowed in Vietnamese syllables ending by /p/, /t/, /k/: sắc mid rising and nặng mid falling.
- The speaker S could be either a man or a woman, one of the two Vietnamese native speakers who recorded each the 144 stimuli (2 first consonants x 6 vowels x 6 last consonants x 2 tones).

Audio recording was carried out in a sound-proof booth, where stimuli written with Vietnamese orthography were displayed one by one on a screen using Matlab. Further audio editing was carried out with Praat.

2.2. Design

A given participant was only exposed to one of the three associations (consonant, speaker, tone), via the selection of stimuli shown to him during exposure. For counterbalancing reasons, there were two opposed rules in each condition (e.g., /f/ occurs with front vowels, tone 1 with front, etc., versus /f/ occurs with back vowels, etc.) Thus, each association implied dividing the final matrix of stimuli in two halves with all other factors counterbalanced, such that one half was legal for half of the participants, and the other half was legal for the others. This ensures that any effects found relate to the *selection* of stimuli heard during exposure (and are not related to freak characteristics of the tokens), since every token is used as training and test, legal and illegal, across participants. Moreover, each participant hears every possible sound, tone, and speaker – they vary only in how these features co-occur during the exposure phase.

Additionally, the 288 CVC-T-S combinations were split into 12 balanced groups. As will be explained below, participants are presented with 120 legal items during exposure, of which 12 are presented again at test; and 24 further legal items are added at test. The division into 12 groups of items allowed us to rotate these test stimuli (e.g., participant one may hear groups 1-10 during exposure, and groups 10-12 during test). This counterbalancing created exactly 72 unique experiences: 3 type of constraints x 2 rules (association with front or back vowels) x 12 grouping conditions.

2.3. Participants

Subsequent analyses focus on 83 speakers of non-tonal languages that could be included (recruitment and exclusionary criteria are given immediately below). These participants had various non-tonal native language backgrounds (English, French, Spanish, Italian, Dutch, German, Turkish, Malagasy, Ukrainian, Arabic, Japanese, Danish, Crucian, Limburgian, Occitan, Romanian), 17 reported to be native multilinguals (all languages being non-tonal), 60 reported to have already received some kind of musical training and 23 reported to be linguists.

The experiment was proposed as a voluntary and non-remunerated 20-minute online experiment on "language and memory" to adults with a tonal or a non-tonal background recruited through:

- Relais d'Information sur les Sciences de la Cognition, a French platform connecting scientists and participants who subscribed to volunteer in experiments in cognitive science: http://expesciences.risc.cnrs.fr/
- LinguistList, an international mailing list addressing linguistics professors and students: http://linguistlist.org/
- word of mouth by e-mails and social networks to potential participants with various language backgrounds
- New York/Queen's craigslist and posters in Paris.

As explained below, we had an independent data quality criterion based on 15 "catch" trials, and self-report of technical issues. The stopping rules for data collection were: 72 tonal and non-tonal speakers without technical difficulties and passing the 13 out of 15 catch trials (72 represented a perfect counterbalancing of all factors); or Dec. 31, 2014.

By Dec. 31, 2014, we had received results from 130 participants, 107 non-tonal and 23 tonal participants. An histogram of the number of correct catch trials confirmed that the mode of this distribution is above 13; this data-quality criterion excluded 15 non-tonal and 4 tonal participants. Exclusion for technical problems concerning sounds on the basis of participants' reports concerned 9 non-tonal and 3 tonal participants. Our final sample therefore consisted of 83 non-tonal and 16 tonal participants. For ease of exposition, we concentrate on the former; please see https://osf.io/kcf6w/ for further information on the tonal speakers.

2.4. Procedure

During the training phase, participants passively listened, one by one and paired with pictures of unknown objects, 120 legal stimuli. This pairing with an image aimed at keeping the participant concentrated on the experiment, without altering his/her listening experience of the stimuli.

Before the experiment started, participants received the instruction to pay close attention to following pictures and sounds, as they would be asked about them later on. They were also asked to click on the center image every time they heard a humming sound: 15 such "catch" trials were randomly presented to check if the participants were on task. Note that we used a nasal sound that does not contain any consonant or vowel, and thus should not interfere with the effects of interest.

At test, they were presented with a 2-alternative forced choice (2AFC) task on pairs of CVC-S-T syllables differing only in the vowel: this minimal change made the two syllables differ in their legal versus illegal status with respect to the exposure phase. They were asked "Which syllable sounds the most familiar (the first or second)?", and they had to click on symbols 1 or 2 on the screen to indicate their response. Each pair of syllables were presented twice, to counterbalance the order of the legal item (which equalizes the number of correct answers for 1 and 2).

The test contained three distinct phases, always presented in the same order:

1. "Memory" phase: The 12 legal items had actually been presented once during the exposure phase; thus, participants should choose it over the illegal one both because it sounds more familiar (since it follows the rule) and because it actually is more familiar (was presented before).

2. "Generalization" phase: None of the 12 legal-illegal pairs had been presented before; a significant preference can only be attributed to extraction and generalization of the underlying rule. There was no break between the memory and generalization phases.

3. "Informed generalization" phase: A screen was shown which explained to the participants the logic of the present study, noting that there are three types of rules related to one of: consonant, speaker, tone. Participants learned that they were only exposed to one of these rules. They were then presented with 12 new legal and illegal items.

We had intended to look at dimensions that are relevant for memory and generalization, as well as explicit recall. However, as will be explained below, the effects of the type of phase cannot be teased apart from the effects of time.

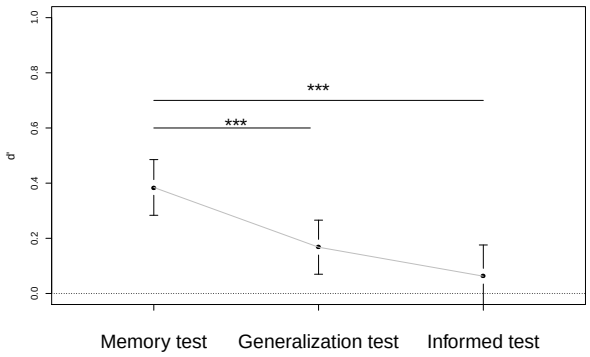

Figure 1: *2AFC performance (d') as a function of Test Phase.*

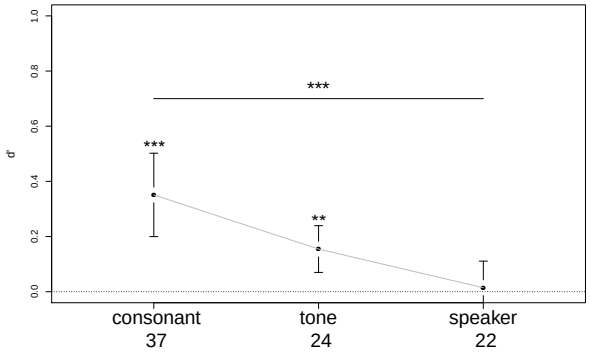

Figure 2: *2AFC performance (d') as a function of Condition. *** p <0.001, ** p <0.01, * p <0.05*

3. Results

Our dependent measure is proportion of correct answers, converted into sensitivity score d' taking into account hits and false alarms for 2AFC tasks [3]. As noted above, there were many different counterbalancing conditions; therefore, we carried out analyses on a balanced sample of 36 participants, and on the total sample of 83 participants. As the same results hold for both analyses, we will only report the latter, which has more power.

Using lme4, we fit a mixed model declaring Condition (consonant-, tone-, speaker-vowel co-occurrence) in interaction with Test Phase (Memory, Generalization, Informed) as fixed effects, as well as participant and the counterbalancing factors (rule and set) as random effects (the precise formula was $lmer(dprime \sim cond * testphase + (1|subj) + (1|rule) + (1|set))$), and reduced models without Condition and Test Phase which we compared against the full model using ANOVA. This revealed a main effect of Condition $\chi^2(6)$ = 21.461 , p = .002, a main effect of Test Phase $\chi^2(6)$ = 37.263, p <.001, and no interaction. Examining Test Phase, there is a significant difference between memory and generalization scores: $t(82)$ = 3.971, p <.001, and memory and informed generalization scores: $t(82)$ = 5.35, p <.001, but not in the scores between the two generalization phases: $t(82)$ = 1.751, p >.05 (Figure 1).

As for the effects of condition on scores averaged over the three phases, there were significant differences in follow-up tests for the consonant and speaker conditions: $t(55.121)$ =

3.832, $p < .001$, but not for the consonant and tone conditions: $t(53.389) = 2.307$, $p > .05$, nor for the speaker and tone conditions: $t(42.758) = 2.266$, $p > .05$. In other words, participants in the tone condition displayed a level of performance that was intermediate between those in the consonant and the speaker conditions. We further assessed whether performance was above chance (d'=0) using a t-test (two-tailed) in the consonant condition: $t(36) = 4.707$, $p < .001$, in the tone condition: $t(23) = 3.777$, $p < .01$, but not in the speaker condition: $t(21) = 0.310$, $p > .05$ (Figure 2).

4. Discussion

This work is motivated by a key theoretical concern: what are the dimensions that are salient when implicitly learning sound patterns? Our results using artificial phonological patterns replicate Onishi et al.'s findings [2] that consonant-vowel co-occurrences are more readily encoded than co-occurrences involving the speaker's voice, and extend them in two directions: by showing that this differential performance can even be found in a purely perceptual experiment, and by testing speakers of a non-tonal language with a co-occurrence between vowel and tone.

The absence of significant discrimination with patterns involving speaker's voice does not imply that language users ignore or normalize for talker identity during early perception. Quite to the contrary, other research demonstrates that listeners incorporate information about talker gender during vowel perception, e.g., [4]. Instead, our results suggest a *relative* learnability difference (not an absolute one): voice information is not *as* relevant *as*, for instance, consonant information when computing, encoding, or retrieving sound co-occurrences, at least in the short time scales observed here. It is likely that, given sufficient exposure and time, such co-occurrences could also be learnt (see also [5]).

The significant results with tone patterns are interesting because they suggest that even though tones are not encoded lexically in the tested languages , they can nevertheless be used to extract co-occurrence patterns (since d' is significantly above zero). Of course, the performance was not numerically as high as for consonant-vowels, which is consistent with the idea that tone is a cue difficult to perceive for non-tonal native speakers [6].

Above chance performance in the tone condition departs from a view whereby lexically contrastive dimensions in one's native language are encoded and non-contrastive dimensions are ignored, based on previous results on infants' learning of allophonic versus phonemic dimensions [7]. It also differs from predictions based on universal capabilities, where potentially linguistically contrastive dimensions are attended to regardless of whether they are used in one's native language, precisely because performance for the tone condition was not significantly better than that found in the speaker condition. Thus, it is possible to interpret this mixed pattern of results as suggestive of an intermediate position, where both cognitive biases and experience modulate performance.

While the strongest positions of both views are thus ruled out, weaker versions of both hypotheses can be postulated. Indeed, the empiricist position could be brought to account for this pattern by enriching it with the assumption that performance can be boosted by attention to any *linguistically* relevant dimension, be it lexically or grammatically relevant. While tone is not used for lexical contrasts, intonation is grammatically meaningful. Since syllables in the stimuli were presented in isolation, participants were free to interpret them as tone or intonation. A follow-up experiment where stimuli are presented in a sentential content could be used to block an intonation interpretation by more clearly anchoring the tone to the syllable.

Similarly, the position that differences in performance are entirely attributable to cognitive biases independent from language experience could be brought to account for the intermediate tone performance by proposing that tones are less perceptually salient than consonants. In our view, this explanation has little traction since perceptual factors should favor the tone-vowel association (e.g., the tone is temporally coextensive with the vowel). Nonetheless, presenting the same exact experiment to tonal speakers could help test this view, since it would predict lower performance for tones than consonants in this population too. (It is too early to draw any conclusions about tonal participants, as they are only 4 in consonant and 3 in tone condition.)

We would like to close by discussing two limitations of the present work, that we will consider in any follow-up of the experiment. First, we did not model the patterns on rules attested in human languages. For example, whereas tone does seem to relate to height (or specifically advanced tongue root, in Slovenian [8]), we know of no language in which tone is related to vowel frontness, and only one of the consonant-vowel conditions was "natural" (alveolar consonants tend to co-occur with front vowels, [9]). Nonetheless, phonetic groundedness is thought to play a minor role in implicit sound pattern learning, which appears to be constrained mainly by formal complexity [1].

A second limitation that should be addressed in future work concerns the salience of different dimensions as a function of the test items. We observed a decrease in performance as the test phase progressed, with a sharp drop between the memory and the generalization phases. The decrease in performance could be due to the use of novel stimuli (which demands generalization) and/or to the fact that participants face the generalization items much later in time, and after a mixed exposure (the generalization phase occurs after having heard 50% illegal items in the pairs of the memory phase). Since we used a constant order, we cannot tease these possibilities apart. Future work should counterbalance the order of test phases, or use a between-subjects design, in order to better understand whether salience of different dimensions is partly dependent on whether generalization and explicit rule knowledge is required.

5. Conclusions

We explored the question of how frequent sound patterns are encoded using implicit learning of artificial phonotactic patterns. Not all co-occurrences are equally salient, which suggests a bias in rule learning, not driven by stimulus perceptibility but by the potential linguistically contrastive nature of the information. Future work could employ variants of this paradigm to tease apart the effects of acoustic salience, linguistic biases, and experience.

6. Acknowledgements

We acknowledge the Agence Nationale pour la Recherche (Grants ANR-14-CE30-0003 MechELex, ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-IDEX-0001-02 PSL*, and ANR-10- LABX-0087 IEC), the European Research Council (Grant ERC-2011-AdG-295810 BOOTPHON), and the Fondation de France.

7. References

[1] E. Moreton and J. Pater, "Structure and substance in artificial-phonology learning, part i: Structure," *Language and linguistics compass*, vol. 6, no. 11, pp. 686–701, 2012.

[2] K. H. Onishi, K. E. Chambers, and C. Fisher, "Learning phonotactic constraints from brief auditory experience," *Cognition*, vol. 83, no. 1, pp. B13–B23, 2002.

[3] J. G. Snodgrass and J. Corwin, "Pragmatics of measuring recognition memory: applications to dementia and amnesia." *Journal of Experimental Psychology: General*, vol. 117, no. 1, p. 34, 1988.

[4] K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory–visual integration of talker gender in vowel perception," *Journal of Phonetics*, vol. 27, no. 4, pp. 359–384, 1999.

[5] D. J. Weiss, C. Gerfen, and A. D. Mitchel, "Speech segmentation in a simulated bilingual environment: A challenge for statistical learning?" *Language Learning and Development*, vol. 5, no. 1, pp. 30–49, 2009.

[6] A. Chen and R. Kager, "The perception of lexical tones and tone sandhi in l2: Success or failure?" in *Congress of Phonetic Sciences XVII proc*, 2011.

[7] A. Seidl, A. Cristià, A. Bernard, and K. H. Onishi, "Allophonic and phonemic contrasts in infants' learning of sound patterns," *Language Learning and Development*, vol. 5, no. 3, pp. 191–202, 2009.

[8] M. Becker and P. Jurgec, "Tone/atr interactions in slovenian," in *Workshop on Segments and Tone, Leiden University*, 2007.

[9] H. Kawasaki-Fukumori, "An acoustical basis for universal phonotactic constraints," *Language and Speech*, vol. 35, no. 1-2, pp. 73–86, 1992.