Title
Functional Load, Perception, and the Learning of Phonological Alternations

Author
Lin, Isabelle

Publication Date
2019

UNIVERSITY OF CALIFORNIA

Los Angeles

Functional Load, Perception, and the Learning of Phonological Alternations

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy
in Linguistics

by

Isabelle Lin

2019

ABSTRACT OF THE DISSERTATION

Functional Load, Perception, and the Learning of Phonological Alternations

by

Isabelle Lin

Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2019
Professor Kie Ross Zuraw, Co-Chair
Professor Megha Sundara, Co-Chair

Languages use combinations of sounds to form words, and some words are only distinguished by a one sound difference (e.g. [d]ank "thank" and [t]ank "tank" in German). Sounds that can induce meaning differences are said to contrast. However, some contrasting sounds can be neutralized in specific positions in the word (in German /d/ becomes [t] at the end of a word, e.g. ra/t/ and ra/d/ both become ra[t]). Within a language, the functional load hypothesis (Jakobson, 1931; Martinet, 1952) states that sounds distinguishing more words (high functional load) are less likely to disappear or merge with other sounds during sound change.

The aim of this dissertation was to examine whether this tendency to preserve highly informative contrasts carries over to the learning of new sound patterns. Additionally, we tested whether learning alternations involving a specific sound pair impacts the perceptual distinctiveness of its members. We used artificial language learning experiments to assess the learnability of alternations involving contrasts with different functional loads in English, and a perceptual discrimination task to assess the perceptual distinctiveness of said contrasts pre- and post- learning.

There are three main results. First, functional load in English predicts the perceptual distinctiveness of contrasts before learning an artificial language. Second, alternations involving high functional load contrasts are learned better, regardless of whether the contrast is neutralized in the artificial language. Finally, learning an artificial language where a contrast is neutralized reduces the perceptual distinctiveness of that contrast post-learning. In summary, we show that phonological contrastiveness in the native language, as well as learning an artificial language, alters perception.

The dissertation of Isabelle Lin is approved.

Bruce P. Hayes

Patricia Keating

Kie Ross Zuraw, Committee Co-Chair

Megha Sundara, Committee Co-Chair

University of California, Los Angeles

2019

Table of Contents

Table of Figures

List of Tables

Acknowledgements

This dissertation would not have been possible without the help and insights of many people, to whom my most heartfelt thanks are due.

First and foremost, I am extremely grateful to my advisors, Megha Sundara and Kie Zuraw, for their mentorship during the entire project. It was their work that I most looked up to when coming to UCLA, and it was an honor to be able to study under their guidance.

Thanks also to my committee members, Patricia Keating and Bruce Hayes for their advice and comments on this work. Through my years at UCLA, they have taught some of the most inspiring classes I have ever attended.

Thanks to all the members of the Phonology seminar for their feedback on this work.

Thanks to my undergraduate research assistants, Sophia Luo and Maree Iligan for their help with running experimental participants.

Many thanks go to my fellow graduate students in the department, who provided such an amazing environment in which to learn how to do research. My cohort and our weekly meetings about syntax homework made it possible to survive first year as one of the students with least knowledge about linguistic theory. I learned much from Michael Lefkowitz, Iara Mantenuto and Scott Gaines, who generously shared their teaching experience.

Thanks to Adam Royer for his unfailing friendship, for the gallery visits, the art parties, the nail painting and the petting of many fluffy cats.

Thanks to Iara Mantenuto for being one of my earliest friends at UCLA, and helping me not feel too lost on this new continent.

How can I thank you, Eleanor Glewwe, for becoming the best friend I never thought I could have? You have been and will always be a role model and a confidante, and the first one to ever encourage me to follow my dearest ambitions. Though words cannot suffice, thank you for the stories and the company, for the late night conversations and conspiring to exceed Messenger's word limit when we could not speak in person. I constantly marvel at being able to call you my friend.

I will never have enough thanks for Olivier Wang for putting up with my being away for five long years, and still providing me with constant personal support. I admire you for all you have accomplished on your side of the ocean, in the meantime. I am infinitely lucky to have met you. I could not have done this without you, nor much else, in fact. Thanks to Æncre for making my apartment a home, for giving me a reason to wake up in the morning and for all the purr therapy.

Thanks to my parents, for making me the person I am, and for having supported me through all the twists and turns in my choice of academic paths. Thanks to my maternal grandparents, who would have been the happiest for me.

This dissertation was supported by a UCLA Dissertation Year Fellowship, and the Student Research Support Committee of the UCLA Department of Linguistics.

All mistakes are my own.

Vita

2014	Research Master in Cognitive Science – Master in Management Ecole Normale Supérieure and HEC School of Management, Paris, France
2015	Graduate Summer Research Mentorship University of California, Los Angeles
2015-2017	Teaching Assistant, Instructor University of California, Los Angeles
2016-2019	Graduate Research Assistant University of California, Los Angeles
2017	M.A., Linguistics University of California, Los Angeles
2017-2018	Richard Diebold Graduate Fellowship University of California, Los Angeles
2017-2018	Andrew Mellon Fellowship of Distinction University of California, Los Angeles

2018	Mellon-EPIC (Excellence in Pedagogy and Innovative Classrooms) Fellowship University of California, Los Angeles
2018-2019	Dissertation Year Fellowship University of California, Los Angeles

Publications

Zuraw, K., Aly, A. M., Lin, I., Royer, A. J. (accepted). Gotta Catch 'Em All: Skills Grading In Undergraduate Linguistics. *Language: Teaching Linguistics*.

Guevara-Rukoz, A., I. Lin, M. Morii, Y. Minagawa. E. Dupoux & S. Peperkamp (2017). Which epenthetic vowel? Phonetic categories versus acoustic detail in perceptual vowel epenthesis. *Journal of the Acoustical Society of America, 142*, EL211-EL217.

de Carvalho, A., Dautriche, I., Lin, I., & Christophe, A. (2017). Phrasal prosody constrains syntactic analysis in toddlers. *Cognition, 163*, 67-79.

Chapter I. Introduction

Languages change over time, and at any point in time, the sounds that contrast within a language can differ. New contrasts can emerge and older contrasts can disappear, or be replaced by those new ones. Within a language, the *functional load hypothesis* (Jakobson, 1931; Martinet, 1952) states that sounds distinguishing more words (*high functional load*) are less likely to disappear or merge with other sounds during sound change. The premise to that hypothesis is that not all sound contrasts contribute the same amount of information in the lexicon. For instance, if English were to lose the contrast between /p/ and /b/, many more words would become homophones, and thus need to be disambiguated by context, than if it were to lose the contrast between /θ/ and /ð/. Additionally, minimal pairs for /p/ and /b/ include rather high frequency words such as *by/buy-pie*, *cup-cub*, while /θ/ and /ð/ distinguishes fewer pairs, where one member is often rather infrequent (e.g. *ether-either, thigh-thy*). As speakers presumably try to preserve distinctions between words during communication, it seems reasonable that more informative contrasts would be better preserved than less informative ones.

In recent years, corpus studies have found evidence for the functional load hypothesis by looking at crosslinguistically attested sound changes (Todd 2012, Wedel et al. 2013).

In this dissertation, I propose that this effect is mediated by a perceptual advantage of high-functional-load contrasts, and that the tendency to preserve highly informative contrasts carries over to the learning of new sound patterns in another language. To test this, I conduct simulations and corpus studies to compare the functional load of various phonological contrasts in English, and use perceptual discrimination and artificial language learning experiments to assess perceptual effects and learnability effects of functional load. These results could inform our understanding of phonological learning in a second language, and the amount of information from the lexicon that might be represented in the phonological grammar.

1. Functional load measures

Phonological contrasts carry different amounts of information in a language. One possible way of quantifying the informativeness of a contrast is functional load (Hockett 1955; Kučera 1963). The higher the functional load of a contrast, the more information it contributes. However, before starting in depth analyses of functional load, we need to examine the different methods of calculating this metric.

In the following sections, we discuss two main methods of computing functional load. One is based on the count of minimal pairs for those segments in the corpus, the other relies on the change of entropy in a corpus upon merging the segments being examined.

a. Minimal pair count

Recent work by Todd (2012) and Wedel et al. (2013) suggest another method for computing functional load, using a count of minimal pairs, possibly relativized to the number of words that contain one of two contrasted segments and could thus potentially be part of a minimal pair (Martin 2017). The minimal pair count method is supported by evidence from corpus analysis showing that pairs of sounds which undergo mergers tend to have fewer minimal pairs in the pre-merger state of the lexicon (Wedel et al. 2013). In the following paragraphs we give a toy example of functional load calculation using minimal pair count.

Let's take for example an 8-word corpus consisting of 4 made-up word types "gib", "gip", "pab" and "tib". Their type frequency in the corpus are in parenthesis in the following list: "gib" (2), "gip" (3), "pab" (1) and "tib" (2). The functional load of the /p/-/b/ contrast by minimal pair count is 1 if we do not relativize to the size of the corpus, since there is only one minimal pair for this contrast, "gib" and "gip". To count the number of possible minimal pairs, the actual pair counts for 1, then "pab" could in theory contrast with non-attested "bab" and "pap" and "tib" could contrast with non-attested "tip" – this brings the number of possible minimal pairs to 4. The minimal pair count relativized to the number of possible minimal pairs is then ¼ = 0.25. If we relativize this measure to the size of the corpus, which contains 4 words, we obtain a functional load value of 0.0625. The frequency of these words does not matter for any of these measures. In this project, we used the minimal pair count relativized to the number of possible minimal pairs and the size of the corpus.

b. Entropy

Surendran and Niyogi (2003) defined one possible measure of the functional load of phonological contrasts using an information-theoretic framework. They defined the functional load of a given contrast in a language L as the change in entropy (Shannon & Weaver 1949) induced by its neutralization, i.e. the reduction of uncertainty in the system by reduction of the phoneme inventory. In their implementation, a language L is represented as a sequence of discrete units x (usually words or syllables), and Shannon's entropy H(L) is computed as:

$$H(L) = -\sum_x p(x) log_2 p(x) \qquad (1)$$

Entropy is then computed for a hypothetical language L' in which a given contrast is inactive. For instance, English without voicing contrasts would be a language in which words like "bat" and "pat" are indistinguishable, but words differing on any other contrast are distinct (e.g. "pat"/"bat" vs. "cat", "mat" …). The functional load of a given contrast is thus defined as the relative difference (in percentage) in entropy between L and L':

$$FL = \frac{H(L) - H(L')}{H(L)} \qquad (2)$$

In this project, we compute functional load according to entropy at the word level. In entropy.

We use again the previous 8-word corpus consisting of 4 made-up word types "gib", "gip", "pab" and "tib". Their type frequency in the corpus are in parenthesis in the following list: "gib" (2), "gip" (3), "pab" (1) and "tib" (2). To calculate the functional load of the /p/-/b/ contrast by entropy change, we first need to calculate the entropy of the initial corpus:

$$H(L) = -(\frac{2}{8} log_2\left(\frac{2}{8}\right) + \frac{3}{8} log_2\left(\frac{3}{8}\right) + \frac{1}{8} log_2\left(\frac{1}{8}\right) + \frac{2}{8} log_2\left(\frac{2}{8}\right)) = 1.91$$

We then consider the same corpus where the /p/-/b/ contrast has neutralized to, say, /p/. The corpus becomes then "gip" (2), "gip" (3), "pap" (1) and "tip" (2). The first two words have become homophonous, and the "gip" form now has a frequency of 5. The entropy in this new system is now:

$$H(L') = -(\frac{5}{8} log_2\left(\frac{5}{8}\right) + \frac{1}{8} log_2\left(\frac{1}{8}\right) + \frac{2}{8} log_2\left(\frac{2}{8}\right)) = 1.30$$

The functional load of the /p/-/b/ contrast by variation in entropy is then:

$$\mathrm{FL}_{\mathrm{entropy}} = \frac{1.91 - 1.30}{1.91} = 0.32$$

If we relativize this measure to the size of the corpus, which contains 4 words, we obtain a functional load value of 0.08. In this project, we used the entropy measure relativized to the size of the corpus.

c. General remarks on both measures

As we can see from the previous sections, both measures rely heavily on the existence of minimal pairs for the contrast at issue. This is directly the case in the minimal pair count measure, but for the entropy measure, if two segments did not distinguish any minimal pairs, entropy would not change at all even if they were to merge – no word would become identical to another. For instance, /h/ and /ŋ/ would have a null functional load in English according to either measure, since their respective distributions in the language (never syllable-final vs. never syllable-initial) makes it impossible for there to exist a minimal pair[1]. This case is extreme, since one could argue that position alone could predict which of the two occurs, but if there are cases where the lack of minimal pairs is due to accidental gaps in the lexicon, this could be more problematic.

It should also be noted that due to the way they are computed, the two measures cannot be compared in terms of numerical value. Functional load by entropy has a minimum value of zero, if merging a contrast does not collapse any minimal pairs (no change in entropy), and a maximum value of one, if all words in the entire lexicon have become identical following the merger (this is not plausible in a natural language if we are considering merging a single phonemic contrast). As to the count of minimal pairs, a raw count will range from 0 to n/2, n being the number of word types in the lexicon (a value of n/2 is implausible, as it would require all words in the lexicon to have a minimal pair twin hinging on the contrast being examined). When relativized to the number of possible minimal pairs, we obtain a measure ranging from 0 to 1 (again, 1 is unlikely as it would mean that every single occurrence of the contrast examined has a minimal pair twin). In the following, we will consider the relativized count, as this takes into account the size of the corpora used to obtain the functional load measures.

Finally note that neither formula for computing functional load is sensitive to position. For instance, for the /p/-/b/ contrast, minimal pairs like “pat”-“bat” contribute equally to this measure as “tab”-

---

[1] According to minimal pair count, /h/ and /ŋ/ have a functional load of 0 since they cannot occur in the same position in a word and thus cannot have minimal pairs. According to entropy, we can also show that the /h/-/ŋ/ has a functional load of 0 using for example a set of words like {hang /hæŋ/, hat /hæt/, cat /kæt/}: the initial entropy of this set is $H(L)=-(p(/\text{hæŋ}/)\log_2 p(/\text{hæŋ}/)+p(/\text{hæt}/)\log_2 p(/\text{hæt}/)+ p(/\text{kæt}/)\log_2 p(/\text{kæt}/)$. If we merge /h/ and /ŋ/ to, say, /h/, the set becomes {/hæh/, /hæt/, /kæt/}, but none of the words become homophonous as a result of this merger, and none of the word frequencies, and thus probabilities, change, so $p(/\text{hæŋ}/) = p(/\text{hæh}/)$. The new system entropy $H(L')=H(L)$, so the /h/-/ŋ/ contrast has a functional load of 0 according to the entropy measure.

"tap" or "tapping"-"tabbing". Of course, one could restrict the environments to be taken into account when computing functional load – e.g. only neutralize a contrast word initially to compute entropy in L', or only counting minimal pairs for the word-initial position. However, when looking at a corpus on the whole, this means that two contrasts could have similar values of functional load with very different distributions in the language. We suspect then that depending on the position considered, the set of contrasts that are most informative in a given language will differ.

2. Effects of functional load

a. Functional load and sound mergers

Studies have found evidence for the role of functional load in sound mergers, as was initially predicted by the functional load hypothesis.

Todd (2012) examined a possible case of ongoing vowel length merger in Te Reo Māori, in which long vowels are becoming more similar in duration to their short counterparts. However, all long vowels have not shortened to the same extent, and the contrast is maintained better for some vowel pairs (e.g. /a/-/aː/) than others (e.g. /i/-/iː/). Using both entropy-based and minimal pair count measures of functional load with many possible choices for additional refinements (including homophones, type vs. token frequency, etc.), he showed that the best preserved contrast (/a/-/aː/) indeed had the highest functional load in the language. For the other vowels, he also found a tendency for higher-functional-load length contrasts to be better preserved.

Extending these results using a crosslinguistic corpus of attested sound changes in a corpus of 8 languages (English, Korean, French, German, Dutch, Slovak, Spanish, and Hong Kong Cantonese), Wedel, Kaplan and Jackson (2013) showed that phoneme pairs undergoing mergers (56 pairs) have significantly lower functional load (as computed by minimal pairs) than other pairs which did not merge (578 pairs). Comparing the minimal pair measure and the entropy measure, they further argue that minimal pairs and phoneme probability[2] taken together predict mergers better than entropy-based functional load. However, they draw this conclusion from comparing superset models including functional load by entropy, count of minimal pairs and phoneme frequency to subset models excluding minimal pairs or phoneme frequency, not both. The entropy measure, as mentioned earlier, depends on the number of minimal pairs for the target contrast, *and* on the frequency of occurrence of the members of the minimal pairs. As such, it should contain some redundant information from both the minimal pair count and the phoneme frequency predictors.

The effects of functional load have also been found in the inventory of alternations in the present state of a language. Silverman (2009) has shown that in Korean, despite the high number of neutralizing phonological alternations, these alternations induce in fact very little homophony. He took this to support the functional load hypothesis in the sense that avoiding homophony preserves lexical information. Building on this finding, Kaplan (2010) used a series of Monte Carlo simulations to show that attested neutralizations in Korean induce less homophony than non-attested ones of similar complexity, even when they eliminate the same number of phonemic contrasts.

b. Functional load and language acquisition

Another area where functional load has been claimed to play a role is language acquisition. Articulatory complexity accounts for a lot in terms of age of acquisition, as the vocal tract develops and motor control of the articulators increases. It is not however the only factor that matters: the same sounds

---

[2] The models reported in Wedel et al. (2013) used the natural logarithm of the token-based probability of the higher-probability member of a phoneme pair, defined as

$$p(\text{phoneme}_\text{i}|\text{corpus}) = \frac{\text{frequency of phoneme}_\text{i}\text{ in corpus}}{\text{Sum of frequencies of all phonemes in corpus}}$$

can be learned at very different mean ages in various languages (even taking into account different articulatory strategies across languages). For instance, the lateral approximant /l/ is relatively complex, and emerges rather late in English (as late as age 6 for adult-like production in onset, medial and coda position, Lin and Demuth 2015) or Québécois French (between 3 and 4 yo, McLeod et al. 2011), but is produced consistently by children learning Quiché Mayan as early as 1;7 to 3;0 yo (Pye, Ingram & List 1987). Pye et al. speculate that the early acquisition of /l/ may result from its high frequency in the language.

More recently, Cychosz (2017) found that functional load as measured by entropy loss correlates positively with earlier acquisition of consonants crosslinguistically. Testing her hypothesis on five typologically unrelated languages (American English, Japanese, Shenzhen Mandarin, Peninsular Spanish and Turkish), her models include phone type frequency and functional load as predictors, and articulatory complexity as a random effect. Overall, she finds that phone frequency is the more robust predictor of age of emergence, but that functional load also plays a significant role. In terms of correlation, functional load actually correlated more strongly with age of emergence than frequency in four out of five languages (English, Japanese, Mandarin and Spanish).

For adult learners, some instructors of English as a second language have turned to functional load as a possible tool to assess which contrasts need more attention. Brown (1988) used phoneme frequency and functional load as measured by minimal pair count to assess the relative importance of contrasts in Received Pronunciation and argue that not all contrasts should receive the same amount of instruction time, since their lexical contributions are not equal. Additionally, he mentioned that speakers with different L1 backgrounds struggle with perceiving and producing different contrasts. Munro and Derwing (2006) found that accentedness and intelligibility ratings of Cantonese-accented English by native speakers of American English were more impacted by errors involving high-functional-contrasts in English (following Brown's assessment (1988), e.g. /l/-/n/) than low-functional-load contrasts (e.g. /d/-/ð/). They pointed therefore to the greater necessity of focusing on high-functional-load contrasts during instruction.

These studies show that functional load has consequences on language learning for both infants and adults.

c. Functional load and perception

Because it assesses the role of phonological contrasts in a given language's lexicon, functional load is a measure of contrastiveness. Trubetzkoy (1939) had proposed that sounds which contrast in all environments in a language might be perceived more clearly than sounds which neutralize in some contexts. There has been experimental evidence that phonological contrastiveness and/or functional load in a speaker's native language impacts the perceptual similarity of sounds.

Boomershine et al. (2008) showed that in a VCV context, adult speakers of English rate [d] and [ð] (contrastive in English) as more different than [d] and [ɾ] (allophonic in English), while Spanish speakers showed the opposite patterns, following the status of contrastiveness of the sound pairs in their native language. This result was first obtained in a rating task. Interestingly, this pattern persists in an AX discrimination task, known for tapping into less phonological and more purely acoustic aspects of perception.

Similarly, Harnsberger (2001)'s study of Malayalam nasals showed that in a forced choice AXB task, native listeners of Malayalam perceive allophonically-related sounds ([n] and [n̟]) as much more similar than phonemic ones ([m] and [ɳ]). In contrast, native speakers of other languages showed different perception patterns for the Malayalam sounds that were not entirely predictable from their languages' inventories and the phonological relationships of these sounds (phonemic or allophonic). This hints that the parallel between contrastiveness and perceptual distinctiveness is not perfect.

Johnson and Babel (2009)'s study comparing Dutch and English speakers also showed that allophonic sounds in Dutch ([s] and [ʃ]) were rated as more similar by Dutch speakers than English speakers. Dutch speakers also rated [θ], which is not part of the Dutch inventory, as more similar to both [s] and [ʃ] than English speakers did. In a speeded discrimination task involving these sounds, the two groups differed in reaction times, though not accuracy. The authors interpreted this as indicative of multiple components of perceptual distinctiveness: the auditory similarity of stimuli, the listener's language's phonetic inventory, and its language-specific patterns of alternation.

Allophones are a case of null functional load: if two segments are in complementary distribution, they cannot form minimal pairs. This is consistent with the conception that they are not contrastive. While the phonemic/allophonic distinction is rather binary[3], functional load provides a more gradient measure of contrastiveness. More recently, Kang (2015) found that functional load (by entropy) was related to the identification of /p/,/t/,/k/ CC clusters in VCCV non-words by English, Korean and Japanese speakers in a 9-alternative forced choice listening task. However, the effect was positive in English and Japanese, but slightly negative in Korean. The author explains this as a language-specific effect: functional load does not predict perceptual accuracy in the same way across languages, but notes that some task effects might be at work, since many participants showed ceiling effects, which weakened the correlations. Bennett, Tang and Sian (2018) found that functional load (by entropy) and contextual predictability influence the perception of stop consonants in Kaqchikel. Their study used an AX discrimination task in noise, replicating Boomershine et al.'s result that these language-specific factors are still active in a task traditionally considered to be purely acoustic.

In this dissertation, we first attempted to correlate functional load and perception, then investigated the influence of functional load on the learning of phonological alternations, and the impact of such

---

[3] though quasi-phonemic contrasts have been documented across languages (Hualde 2004, Scobbie & Stuart-Smith 2008)

learning on perception. Our questions were as follows: (1) does the perceptual accuracy of a pair of segments increase with their functional load? (2) is it easier or more difficult to learn alternations involving sounds from high functional load contrast? (3) does learning such alternations impact the perceptual accuracy of these sounds? We used an artificial grammar learning task to test the learning of alternations. Following the studies cited above, we also chose to use an AX discrimination task in noise to assess the perceptual effects under investigation.

3. Learning alternations

When learning a second language (L2), a speaker has to acquire a new inventory of sounds along with the patterns that govern their distribution. However, they inevitably bring along a significant amount of prior knowledge from their native language (L1). In phonetics and phonology, interference effects between L1 and L2 have been found along multiple dimensions. For instance, segments that are non-distinctive in L1 can be harder to perceive as distinct in L2 (Best, 1994), and can also be harder to acquire (Iverson et al., 2003).

In our search for synchronic effects of functional load, we wondered whether interference effects can be predicted on the basis of the informativeness of a specific dimension in L1 and L2, or an artificial L2.

a. Learning alternations in artificial learning grammar experiments

Previous studies have shown that a relatively short period of exposure in an experimental setting is enough for adults and even infants to learn phonological alternations in an artificial language (Peperkamp & Dupoux, 2007; Skoruppa, Lambrechts & Peperkamp 2012; White & Sundara 2014, Moreton & Pater, 2012a, 2012b for an overview). This learning, however, is impacted by a number of biases.

One such bias is that learners are sensitive to phonetic similarity when acquiring phonological patterns (Steriade, 2001). Skoruppa et al. (2012) compared the learning of 6 artificial languages with the same segmental inventory but different alternation patterns, and found that phonetic distance between the alternating sounds affects gradiently adult learners' performance at acquiring said patterns. Alternations involving minimal changes (one feature difference, e.g. [p]-[t], place) were always easier to learn than those involving more changes (two or three feature difference). Looking at saltatory alternations (e.g., /p/ becomes [v] while /b/ remains [b]), White (2014) found that this typologically rarer pattern is also more difficult to learn for adults than non-saltatory alternations (e.g., /b/ becomes [v] while /p/ remains [p]), even when participants were given explicit training. [p] and [v] differ by both voicing and manner features, while [b] and [v] differ only by manner. When participants were shown that the less similar sounds ([p] and [v]) alternated in a language, they tended to assume that the more similar ([b] and [v]) would also alternate, despite evidence to the contrary. He uses this finding as evidence that participants expect alternations between perceptually similar sounds to be more likely.

This bias for learning alternations involving similar sounds is also found for 12-month-old infants (White & Sundara 2014). Infants exposed to alternations between less similar sounds ([p]-[v]) generalized the alternation to more similar sounds ([b]-[v]), but infants exposed to alternations between more similar sounds did not generalize to less similar sounds.

These results show that phonetic similarity affect phonological learning. Interestingly, this is almost a mirror image of the results from the previous section, where phonological contrastiveness in a language impacts its speakers' perception. By measuring the contribution of phonological contrasts to distinguishing words in the lexicon, functional load is a measure of contrastiveness that is more gradient than a binary phonemic/allophonic distinction. In this dissertation, we will be examining whether functional load impacts phonological learning, and/or perceptual distinctiveness of sound contrasts.

b. Learning neutralization

In this project, we investigated the learning of a specific type of alternation, neutralization. Neutralization is a phonological process by which distinct phonemes are realized as identical phones in specific environments. For instance, even though English and German both have a /t/ and a /d/ sound, and they contrast in both languages ([d]ime and [t]ime are different words in English, and [d]ank "thank" and [t]ank "tank" are different words in German), in German /d/ becomes [t] at the end of a word (ra/t/ "advice" and ra/d/ "wheel" are both realized as ra[t]) but not in English (/pad/ and/pat/ become pa[d] and pa[t]). The /t/-/d/ contrast is said to be neutralized word-finally in German.

Recent results by Yin and White (2018) have shown that when participants are exposed to equal amounts of evidence for neutralizing (e.g. /t/ and /tʃ/ both become [tʃ] before a plural suffix [i] in the artificial language) and non-neutralizing alternations (e.g. /s/ becomes [ʃ], /z/ becomes [ʒ]; before [i], and [ʃ] and [ʒ] do not appear elsewhere in the data), learners acquire neutralizing and non-neutralizing patterns to the same degree, except if neutralization induces homophony in the artificial language (e.g. if /tusut/ and /tusutʃ/, two different words, would both become [tusutʃi] in their plural form). This suggests that participants track lexical information in the language they are learning while making hypotheses about its phonological system.

If we consider the fact that every participant walks into an artificial grammar learning experiment with all their experience from their L1, there might be something to be said for the importance of the sounds from L1 that are involved in an alternation (neutralizing or not) in the artificial language. The question we intended to address was then: how well do speakers learn alternations that involve sounds from more or less informative contrasts in their native language?

We propose that the learnability of an alternation could be predictable based on the informativeness a specific contrast, in L1 and/or in the artificial language. Furthermore, successfully learning alternations entails learning a different phonemic status for the sounds involved (allophonic or phonemic) in the artificial language. If phonological status influences perception, we should observe a change in perception following the artificial grammar learning experiment.

4. Structure of the dissertation

The objective of this dissertation was to look for synchronic evidence of the functional load hypothesis. We conducted this search following two directions: (1) looking for effects of functional load in the current state of a language, and (2) looking for effects of functional load in the learning of an artificial language. The structure of the dissertation is as follows.

We first conduct a series of simulations on toy corpora to examine how similarly the two measures of functional load, the variation in entropy and the count of minimal pairs, behave in particular situations. This allows us to explore conceptual differences between the two main methods of computing functional load. We then turn to corpus studies on real languages to see whether these differences matter as much in real languages (chapter 2). We then examined the relationship between functional load and perception. To this end, we attempted to correlate measures of functional load in English and French to pre-existing data on the perceptual distinctiveness of English consonants (chapter 3). As some uncertainty remained from these results, we collected more data using a perceptual discrimination experiment and an artificial grammar learning experiment; we examined the effect of functional load on the learning of phonological alternations, and the effect of such learning on perception (chapter 4). Finally, we summarize our findings and discuss directions for future research (chapter 5).

Chapter II. Simulations and corpus studies

In order to compare measures of functional load by entropy and by minimal pair count, we conducted a series of simulations and corpus studies. As described in the previous chapter, both measures have been shown to correlate with observed data crosslinguistically (historical mergers, age of acquisition), but how similar are they? We began by comparing them in a series of small toy corpora, then examined their relationship to perception in English.

1. Toy corpora simulations

First, we conducted some simulations using toy corpora with a limited segment inventory to compare how both measures fared in various scenarios. In particular, we were testing for conditions in which the two measures would be most similar or dissimilar.

a. Method

We wrote a Python script to generate small corpora from a limited set of segments, with random number of observations and various constraints on word shape and segment distribution. For each set of conditions on word shape and segment distribution, we ran 1000 iterations of the corpus-creating process and calculated, each time, functional load based on entropy and based on minimal pair count. For entropy, we use word type frequency, and for minimal pairs, the count of minimal pairs relativized to the number of possible minimal pairs. Both measures were computed using Phonological Corpus Tools (Hall et al., 2018), and were relativized to the size of the corpus.

- *Segment inventory*

All words were made up from a random combination of the following segments (Table 1), with various constraints on segment distribution. For ease of reading, we used the combined letter strings as phonetic transcriptions (i.e. tup = [tup]).

Table 1. *Segment inventory for simulations with toy corpora*

Consonants	Vowels
p	a
b	i
t	u
d	

We chose this very limited inventory to be better able to engineer specific cases to compare functional load measures. Calculations with more natural inputs would be conducted with real language corpora, so realism was not our aim here.

- *Lexicon size and frequencies*

Each lexicon contained 15 word types. For each word, we generated a random number of observations between 1 and 25 (uniform distribution). Each toy corpus then contained between 25 and 25 x 15 = 375 observations. This allowed us to simulate token frequency for each word. Depending on the word-generating conditions, this also allowed us to generate n-tuples of homophones with different token frequencies for each instance of the same wordform (e.g. tup_1 observed 12 times, tup_2 observed 39 times in a corpus with 270 observations).

- *Word shape and segment distribution*

We created corpora with specific conditions on wordforms to test the variation they would induce on the different measures of functional load. For simplicity, all words were 3 segments long.

Our first condition consisted of entirely randomly selected segments (XXX words). Any segment was allowed to occur in any position, and segments were allowed to reoccur in a word (e.g. ppp, iat, taa, idd, aaa…). This allowed for $7^3 = 343$ possible wordforms. The probability of having one or more pairs of homophones in a 15-word types corpus was 0.264[4]. That of having at least one minimal pair (for any pair of segments) in one corpus was 0.997[5]. This provided a baseline to compare the functional load measures in a case where a given contrast could operate in any position in the word.

Our next condition consisted of Consonant-Vowel-Consonant words, where segment distribution was limited by segment type. Segments were allowed to reoccur in a word (e.g. tut, pap…). This allowed for 4*3*4 = 48 possible wordforms. The probability of having one or more pairs of homophones in a 15-word types corpus was 0.888. Unless otherwise constrained, that of having at least one minimal pair was 1. We used this case to simulate basic restrictions on word-shape.

A third simulation involved once again Consonant-Vowel-Consonant words, but this time the word-final consonant could only be voiceless (p or t). This allowed for 4*3*2 = 24 possible wordforms. The

---

[4] With 15 words, $\frac{15*14}{2} = 105$ pairs of words need to be evaluated as to whether they are identical. The probability of 2 words being different becomes $1 - \frac{1}{343}$, and that of there being no homophones in the corpus, $\left(1 - \frac{1}{343}\right)^{105} = \left(\frac{342}{343}\right)^{105}$. The probability of having at least one pair of homophones is therefore $1 - \left(\frac{342}{343}\right)^{105} \approx 0.264$

[5] If we define a minimal pair as two words differing by exactly one segment, each word has 18 possible candidates for forming a minimal pair (only one position can differ per word, 6 different segments are available for each position, and the difference could be on any of the 3 segments). For any given word, there are thus 343-18 words that do not form a minimal pair with it. The probability for there to be at least one minimal pair in a 15-word corpus of the sort becomes $1 - \left(\frac{343-18}{343}\right)^{105} = 0.997$

probability of having one or more pairs of homophones in a 15-word types corpus was 0.987. Unless otherwise constrained, that of having at least one minimal pair was also 1. In this system, voicing is contrastive, but only word-initially. We used this restriction to simulate cases in which contrast is limited to some positions.

We also consider cases (with any segment structure), where there are no minimal pairs, and cases where minimal pairs only involve some segment positions.

b. Results

We report functional load comparisons for simulations neutralizing the voicing contrasts on stops. Neutralizing place features on stops or height, frontness/backness on vowels yield similar results in these simulations.

- *XXX – no positional constraints*

  In 1000 small corpora with no positional constraints on any of the possible segments, we examined the functional load of the voicing contrast in stops. We plotted the entropy and minimal pair count measures against the other (Figure 1) to visualize how they correlated.

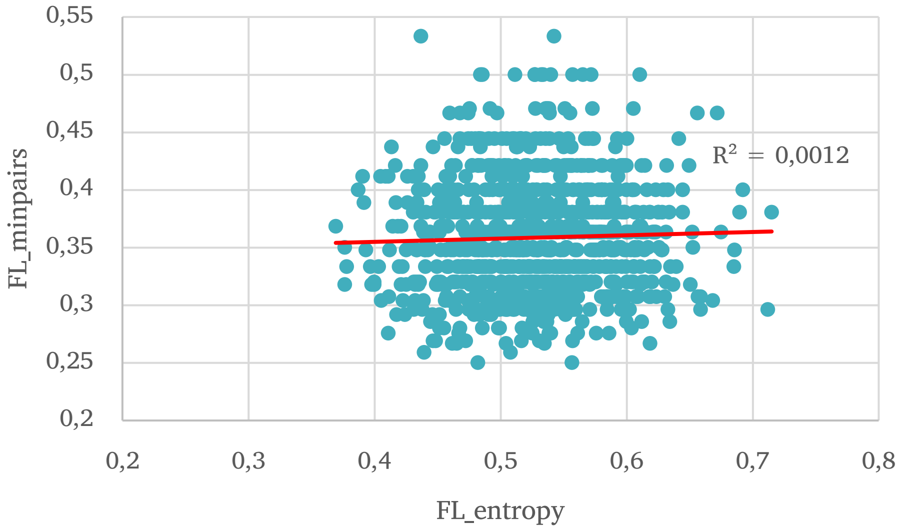


Figure 1. *Functional load measures for the voicing contrast in 1000 corpora made up of words with no positional restrictions on segments*

When no positional constraints at all were placed on the possible segments in the language, we found no significant correlation[6] between minimal pair and entropy measures ($r_s = 0.04$, $p = 0.19$). This case was the largest possible amount of lexical variability. Because of the small size of each corpus (15 words), even though it was likely to have at least one minimal pair for a given pair of segments, the likelihood that this pair differed in voicing only would still be rather low. With very few minimal pairs, the entropy measure would be strongly affected by the frequency of very few words, which could happen to be extremely high or low. Conversely, since the minimal pair count measure is unaffected by frequency, the two measures could yield very different outcomes.

---

[6] According to a Shapiro-Wilk test of normality, functional load values as measured by entropy were normally distributed ($W = 0.99$, $p = 0.52$), but functional load as measured by minimal pair counts was not ($W = 0.97$, $p < 0.001$). To examine the relationship between the two variables, we henceforth use the Spearman correlation test, which does not carry any assumptions about the distribution of the data.

- *CVC syllables*

With an additional condition on word structure, we do find a significant correlation between the minimal pair and the entropy measures, but the coefficient of correlation remains poor ($r_s = 0.17$, $p < 0.001$, Figure 2). The same issue reappears as the previous case – however, this time, since consonants can only occur in two positions, there is a greater likelihood of finding minimal pairs differing in voicing.

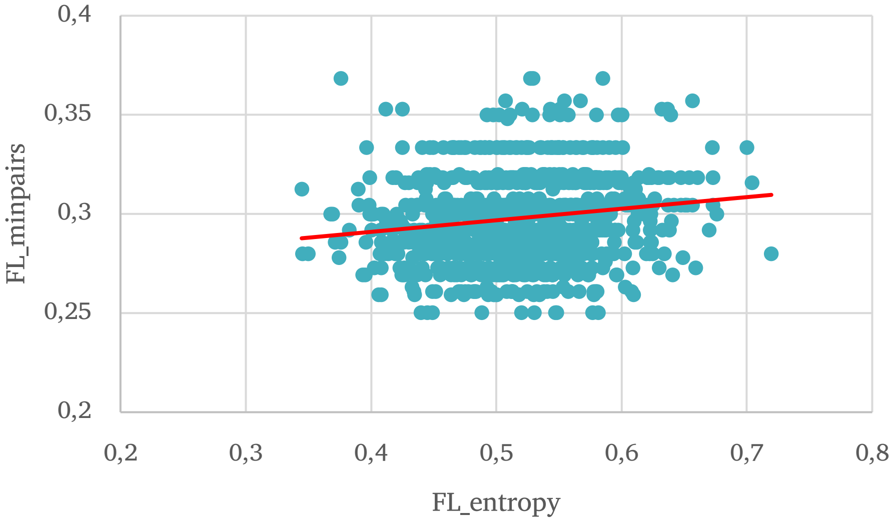


Figure 2*. Functional load measures for the voicing contrast in 1000 corpora made up of CVC words*

- *CVC syllables, positional voicing contrast*

The improved correlation between functional load measures with more constrained wordforms remains when we constrain furthermore the voicing feature on the last segment of the words ($r_s = 0.24$, $p < 0.001$, Figure 3). With fewer possible words, and voicing contrasts occurring only word-initially, making minimal pairs more likely in any given corpus, and for those minimal pairs to involve a voicing-only difference. The entropy measure takes into account the frequency of multiple words, and thus is less sensitive to extreme values. We expect thus that natural languages, with more constrained segment distributions, will show a better correlation between functional load measures.

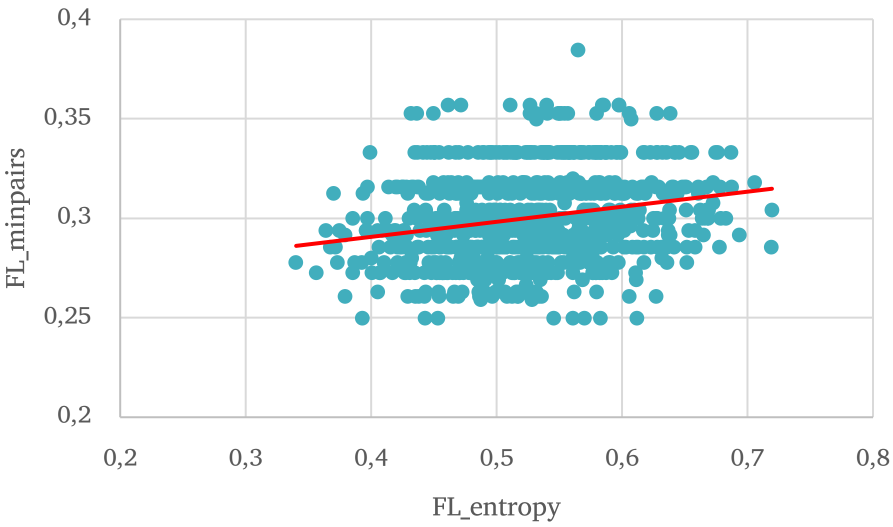


Figure 3*. Functional load measures for the voicing contrast in 1000 corpora made up of CVC words ending in voiceless consonants*

- *No minimal pairs*

No matter the size of the corpus or the shape of the words, having minimal pairs for a given contrast is crucial for either measure to result in a non-null value. Therefore, if we constrain our corpora so that no minimal pairs are present even though two segments share the same distribution, we obtain cases where functional load is numerically null. Consider for instance the following 15 CVC words:

tap dat tup pud tib bud dip bap
bip pit bup dut tab tid pad

The distribution for [d] and [t] completely overlaps and both appear in a good number of words, but there are no minimal pairs for these two segments. If we use minimal pair count to represent functional load, regardless of whether we relativize the result to the size of the corpus or the number of possible minimal pairs, we obtain a value of 0 for the t-d contrast. If we use an entropy measure with type frequency, the initial entropy of the system is $H_0 = -15 * \left(\frac{1}{15}\log_2\frac{1}{15}\right) = \log_2 15$: to compute the entropy after neutralizing the t-d contrast, H_1, we neutralize all t and d segments to the segment X, our corpus becomes:

Xap XaX Xup puX Xib buX Xip bap
bip piX bup XuX Xab XiX paX

where each word retains its initial frequency. After this process, no two words have actually merged: the entropy of the system did not change. The functional load of the t-d contrast as measured by change in entropy is therefore also null.

We note however that there are cases of corpora with no minimal pairs for a contrast where an entropy measure could still yield a non-null result. If we change one word to the previous corpus:

tap dat tup pud tib bud ***dit*** bap
bip pit bup dut tab tid pad

The count of minimal pairs is still null, and so is functional load measured that way. However, if we neutralize t and d, we obtain the following set:

Xap XaX Xup puX Xib buX XiX bap
bip piX bup XuX Xab XiX paX

Neutralizing the contrast will actually collapse two words this time, resulting in 14 word types instead of 15, and $H_1 = -14 * \left(\frac{1}{14}\log_2\frac{1}{14}\right) = \log_2 14 < H_0$. Functional load as measured by entropy will therefore be non-null, $FL_{entropy\_type} = \frac{\log_2 15 - \log_2 14}{\log_2 15}$ = 0.025, though it is still likely to be fairly low – it would compare to that of two segments which only appear in 2 words that form a minimal pair, e.g. f and v in the following corpus:

tap dat tup pud tib bud dit ***baf***
bav pit bup dut tab tid pad

With 15 word types, H_0 is the same as previously: $H_0 = \log_2 15$. If we neutralize f and v to Y, we obtain again 14 word types:

tap dat tup pud tib bud dit ***baY***
baY pit bup dut tab tid pad

And $H_1 = \log_2 14$, $FL_{entropy\_type} = \frac{\log_2 15 - \log_2 14}{\log_2 15}$ = 0.025. Intuitively, the t-d contrast is much more active in this lexicon than the f-v contrast[7]. However, neither measure captures intuitions for contrasts that do not have minimal pairs.

- *Same number of minimal pairs, positionally constrained vs. unconstrained contrast*

An interesting observation arises when we compare cases of word lists with the same number of minimal pairs, but where the contrast is allowed at any point in the word or only word-finally. Despite distributions being very different, both functional load measures will yield the same value in the two

---

[7] The f-v minimal pair could involve words with extremely high token frequency, but with the much wider distribution of t-d, it would be unlikely that they do not also appear in high frequency wordforms.

languages. It seems intuitively that functional load measures could fail to capture some information on contrastiveness in such a case. Let's consider the following example in Table 2:

***Table 2**. Examples for same functional load in different cases*

	Language A	Language B
contrast 1	same positional restrictions (or lack thereof)	positionally constrained
contrast 2		in all positions

Within each of the 2 languages, contrasts 1 and 2 have the same functional load (regardless of which measure we use). In language A, since both contrasts occur in exactly the same positions, it seems natural to conclude that they play roles of similar importance in the language. In language B, however, contrast 1 contributes the same amount of contrastiveness as contrast 2 while occurring in more limited positions. It could be argued in a sense then that contrast 1 needs to be more active than contrast 2 in order to achieve this[8].

- *Homophones*

For this project, we chose not to distinguish homophones for the computing of functional load. Since homophones are distinguished by context most of the time (syntactic position of semantic content), this is not something that can quite be addressed with contrastiveness at a phonological level. In fact,

---

[8] In natural languages, a case where only one sound in a contrasting pair of sounds appears in a certain position can arise from at least two different situations. This positional restriction could be due to an accidental gap, but also to a neutralizing process. In the case of an accidental gap, there never was an opportunity for contrast in that specific position. Meanwhile, in the case of a neutralizing process, the contrast that is active elsewhere in the language is ***lost*** in that position. This could produce yet another distinction in how important the contrast is, but we will not explore this further in this project.

as functional load is computed using only the phonological form of words, it cannot take context into account at all. This means that for a same number minimal pairs (which affects both minimal pair count and entropy), we cannot take into account the fact that some pairs are more informative than others. For instance, a minimal pair whose members belong to different parts of speech (e.g. English *bad* and *bat*) is less crucial than one whose members both belong to the same part of speech (e.g. *cat* and *rat*), as the part of speech of a following word can often be inferred from preceding context. More detailed semantic properties also help deciding which member of a pair is being used. This is beyond the scope of this current project.

In the previous simulations, we found that more constrained word shapes led to a closer correlation between the minimal pair count and entropy measures. Table 3 summarizes the results in the conditions we examined.

***Table 3**. Simulation results for different constraints on the simulated lexicon*

	XXX: no positional constraints		CVC syllables		CVC syllables, positional voicing contrast	
	entropy	minimal pairs	entropy	minimal pairs	entropy	minimal pairs
$\mu_{\text{functional load}}$	0.53	0.36	0.52	0.30	2.17	0.90
$\sigma_{\text{functional load}}$	0.06	0.05	0.06	0.02	0.28	0.24
r_s	0.04		0.17		0.24	
p-value	0.19		< 0.001		< 0.001	
$\mu_{\text{minimal pairs}}$	0.81		4.20		2.53	
$\sigma_{\text{minimal pairs}}$	0.85		1.22		1.07	
$\mu_{\text{homophones}}$[9]	0.28		2.09		3.76	
$\sigma_{\text{homophones}}$	0.50		1.21		1.33	

---

[9] For our simulations, we did not distinguished homophones. This means that in a lexicon of 15 words, if there were 3 homophones (e.g. *bap*, *bap*, *bap*), the lexicon was considered to have 13 wordforms, and the frequencies of the 3 homophones would be summed to correspond to a single wordform (*bap*). If we did not merge homophones, this would overestimate the number of minimal pairs: a single word (e.g. *bab*) could form a minimal pairs with each of the 3 homophones, thus contributing 3 minimal pairs using identical wordforms.

Though correlations for CVC syllables and CVC syllables with a positional voicing contrast were significant, they still had very low coefficients. Overall, we showed with these simulations that the two measures of functional load could mathematically be very different, but became more similar as more constraints were imposed on the structure of words. These constraints increase the probability of occurrence of minimal pairs for the contrast we examined (voicing in stops). Therefore, the weakness of the correlations we obtained might have been due to our toy corpora being too simple. We now turn to corpora in real languages, with many more restrictions on how segments are allowed to pattern, to test whether this is the case.

5. Corpus study: English

a. Methods

To examine the functional load of various contrasts in English, we used data from the IPhOD corpus (Irvine Phonotactic Online Dictionary, Vaden, Halpin & Hickok, 2009), which contains an American English phonetic transcription from the Carnegie-Mellon Pronouncing Dictionary (Weide, 1994), and written word frequencies from the SUBTLEXus database (Brysbaert & New, 2009). Functional load by entropy was calculated using word type frequencies, and minimal pair count was relativized to the number of possible minimal pairs. Both measures were relativized to the size of the corpus. Calculations were performed using Phonological Corpus Tools (Hall et al., 2018).

As previously, we examined the correlation between the two measures, using consonantal contrasts: these contrasts in particular will be examined in further computations. Additionally, we tested whether the two sets of measures come from comparable distributions. If they do not, this would mean that a significant number of contrasts rank differently in terms of functional load according to the two measures. We take a closer look at some specific segment contrasts that exhibit differences between the two measures. We will later exploit these differences in a series of experiments to compare the two measures.

b. Results

In a real language corpus, the two measures were much closer correlated than in previous simulations. Figure 4 shows the entropy measure plotted against the minimal pair count after excluding outliers where either of the two measures fell more than two standard deviations from the mean in each measure ($\mu_{\text{FL\_entropy}} = 0.0006$, $\sigma_{\text{FL\_entropy}} = 0.0006$; $\mu_{\text{FL\_minpairs}} = 0.01$, $\sigma_{\text{FL\_minpairs}} = 0.01$). This criterion excluded 7 out of 120 contrasts. Neither measure followed a normal distribution (Shapiro-Wilk test for normality, $W_{entropy} = 0.88$, $p_{entropy} < 0.001$; $W_{minpairs} = 0.96$, $p_{minpairs} < 0.001$), so we tested for correlation with a Spearman's test. We found that in a real language corpus, the two measures are highly correlated ($r_s = 0.84$, $p < 0.001$).

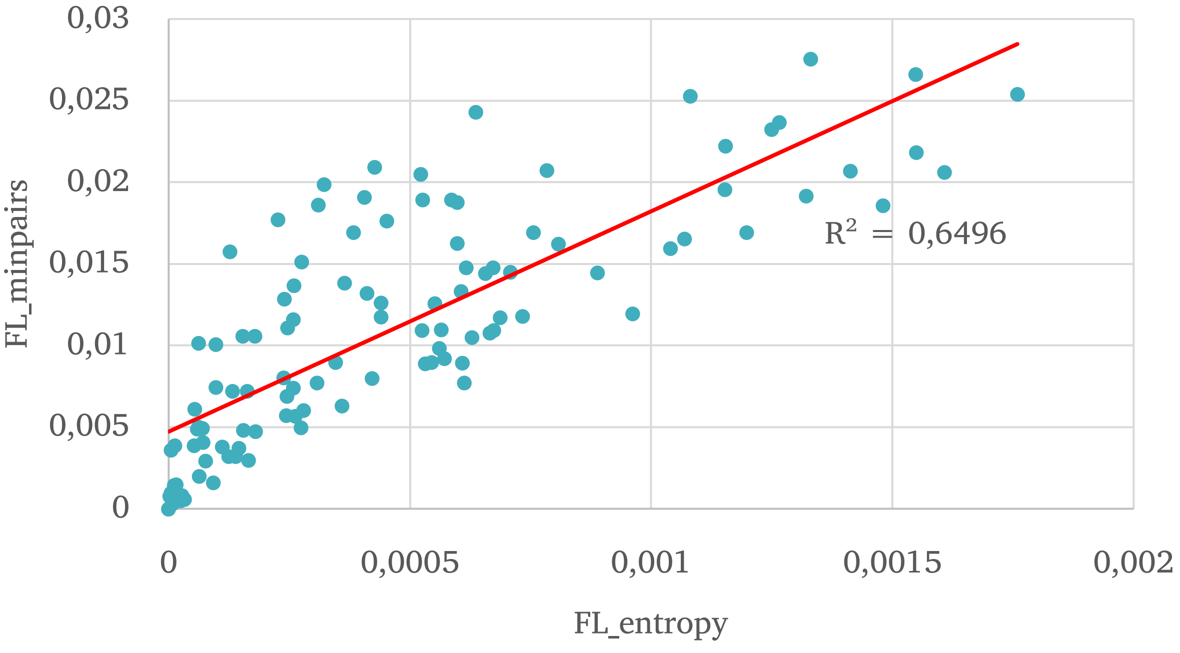


***Figure 4**. Functional load measures for consonantal contrasts in English*

Despite this close correlation, however, the two measures remain significantly distinct in terms of distribution (two-sample Kolmogorov-Smirnov test on centered functional values in the two measures, $D = 0.56$, $p < 0.001$). This means that minimal pairs and entropy make slightly different predictions about which pairs of sounds have higher functional loads. Figure 5 shows the functional load according to entropy and according to minimal pairs for some selected contrasts in English (see full list in Appendix I). For instance, a minimal pair count favors the /p-b/ contrast over the /t-d/ contrast, and the opposite is true for the entropy measure. This effect is driven by word frequency –there are more minimal pairs relying on the /p-b/ contrast in the lexicon, but these words are overall less frequent that than those in minimal pairs hinging on the /t-d/ contrast.

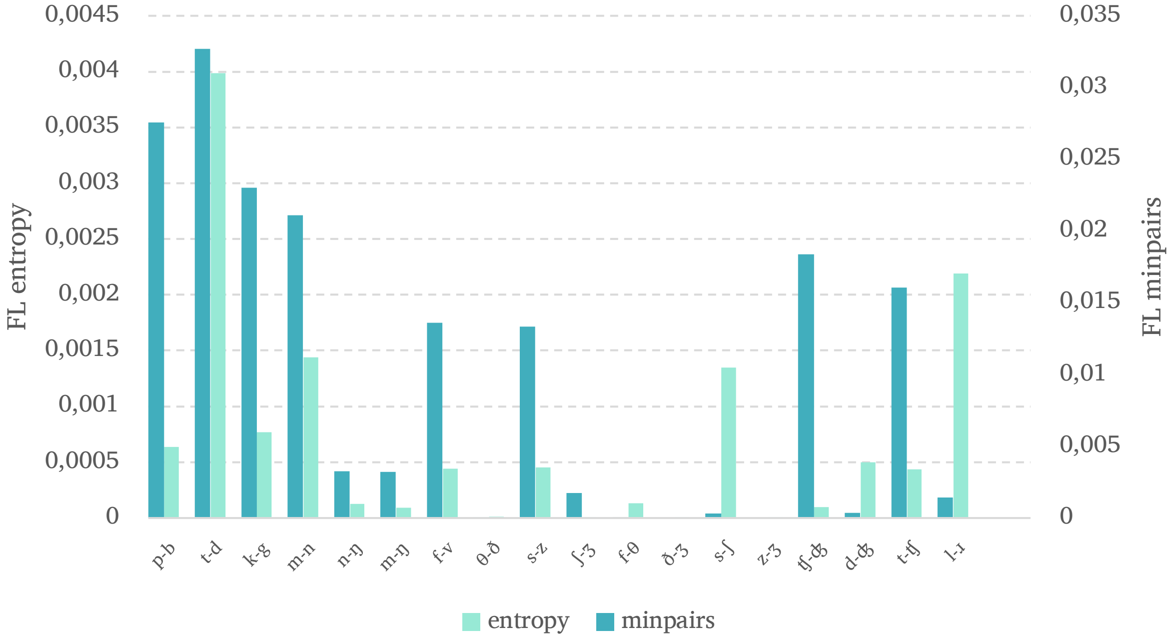


***Figure 5**. Functional load comparisons for English*

We also computed functional load in word initial and word final position for a subset of the previously examined contrasts. Figure 6 shows these results. Most contrasts are more informative word-initially. This is consistent with the cross-linguistic tendency for onsets to be more perceptually salient, and more faithfully preserved (Beckman, 1998) . We note that this is not always the case (e.g. /m/ and /n/ word finally according to the entropy measure).

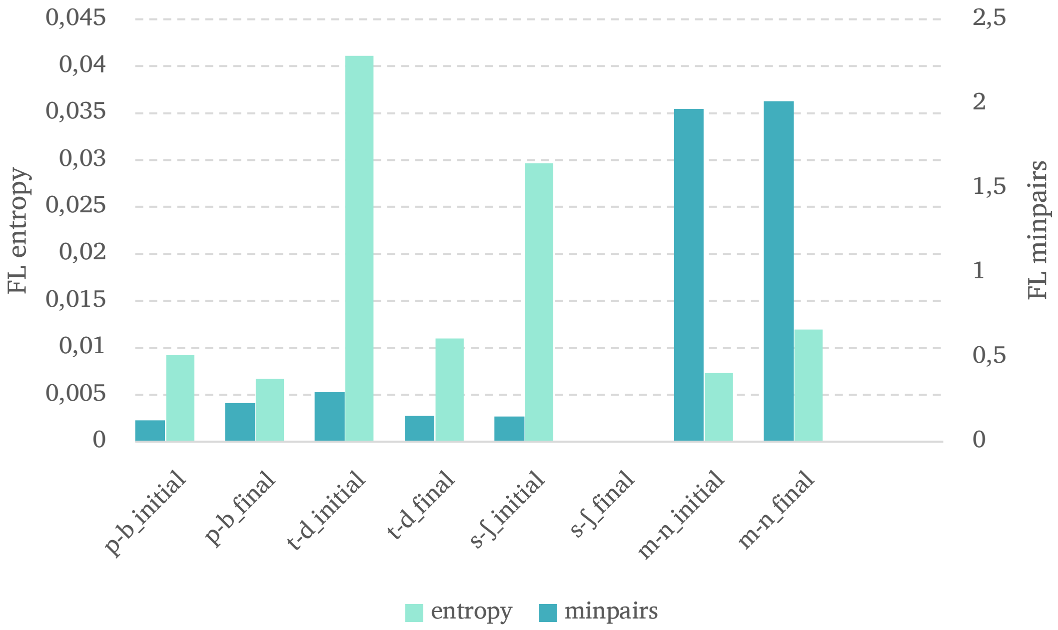


Figure 6*. Functional load comparisons for English - word-initial and word-final positions*

6. Corpus study: French

a. Methods

We conducted a similar analysis for French, using data from the Lexique corpus (New et al., 2001). This corpus contains word frequencies according to various sources, including movies and several types of print. We used the frequencies from movies, as this should provide a better approximation of natural speech. As before, functional load was computed by entropy using word types, and minimal pair counts relativized to possible minimal pairs. Both measures were relativized to the size of the corpus.

b. Results

In French just as in English, the two measures were quite closely correlated. Figure 7 shows the entropy measure plotted against the minimal pair count. Neither measure followed a normal distribution (Shapiro-Wilk test for normality, $W_{entropy} = 0.82$, $p_{entropy} < 0.001$; $W_{minpairs} = 0.96$, $p_{minpairs} = 0.03$), so we tested for correlation with a Spearman's test ($r_s = 0.92$, $p < 0.001$).

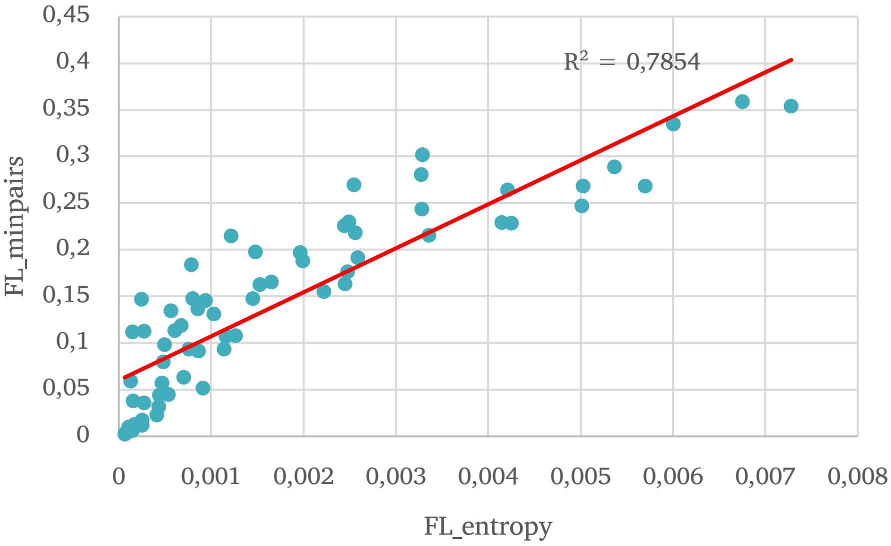


Figure 7*. Functional load measures for consonantal contrasts in French*

Just like the English case, the two measures remain significantly distinct in terms of distribution (two-sample Kolmogorov-Smirnov test on centered functional values in the two measures, $D = 0.5$, $p < 0.001$). Again, this means that minimal pairs and entropy make slightly different predictions about which pairs of sounds have higher functional loads. Figure 8 shows functional load measures for some selected contrasts. Just as for English, the two measures of functional do not quite align as to which contrasts are more informative.

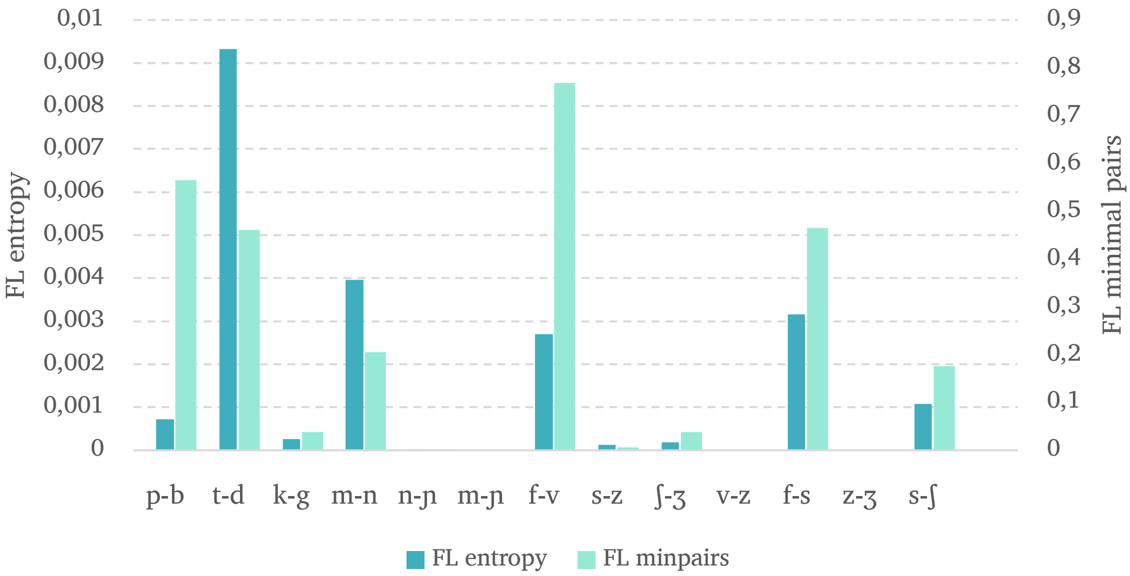


Figure 8*. Functional load comparisons for French*

We examined the same positional contrasts we compared earlier in English. There is also a positional difference for these contrasts in French - it is less clear, however, that word initial contrasts carry more functional load than word final ones (Figure 9).

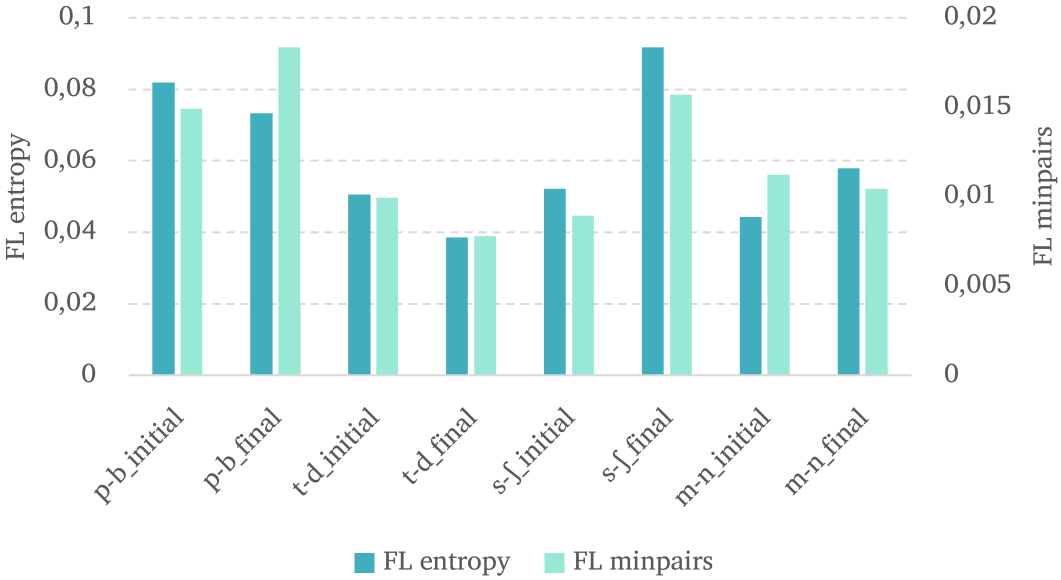


Figure 9*. Functional load comparisons for French - word-initial and word-final positions*

7. Chapter conclusion

In this chapter, we first conducted a series of simulations on toy languages constructed to test the mathematical properties of the two measures of functional load. While we found that have the mathematical potential of being very different, we proceeded to show that they are in fact well correlated in real language corpora. Despite this correlation, they exhibit significantly different distributions in both English and French: the two measures do not rank sound contrasts in the same order, and the same contrast can have a high functional load according to one measure, and a low one according to the other. In the next chapters, we examine the relationship between functional load and other factors, such as contrastiveness and perception. Specifically, we attempt to test for synchronic effects of the functional load hypothesis and to compare the two measures of functional load.

Chapter III. Functional load, perception, and contrastiveness

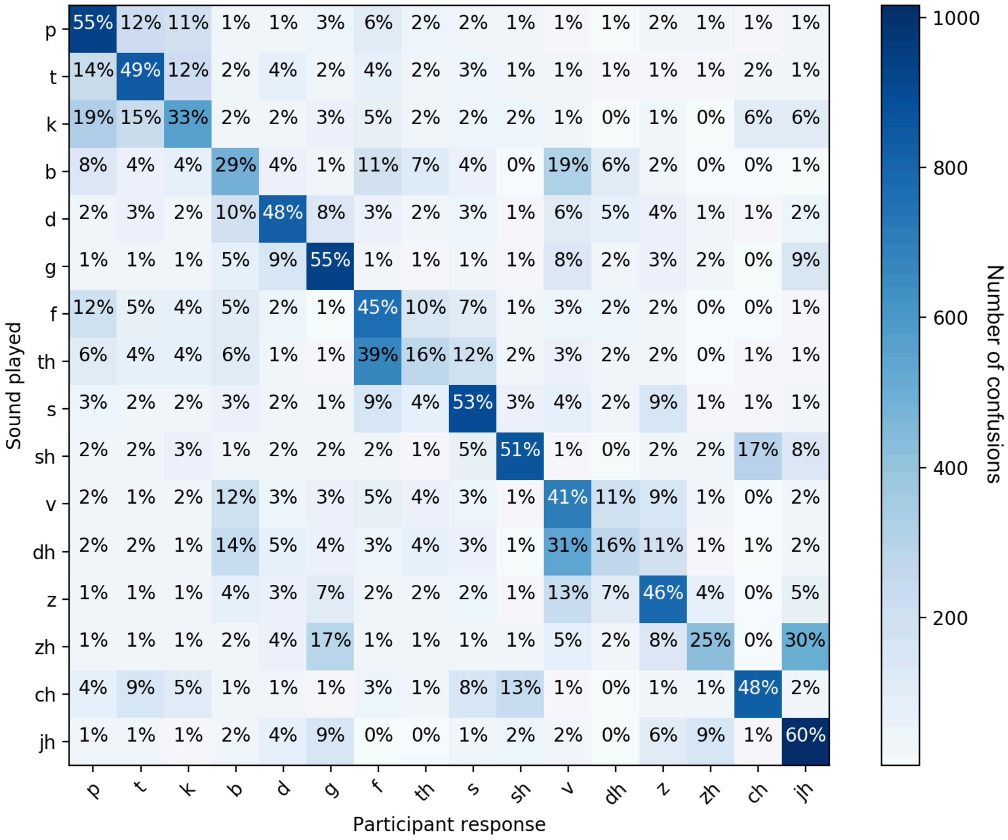
Previous studies have shown that functional load plays a role separate from phoneme frequency and distribution both in how they affect the likelihood of mergers (Todd, 2012; Wedel et al., 2013) and the age of acquisition of a contrast (Van Severen et al., 2013; Stokes & Surendran, 2005; [3]Cychosz, 2017). If distributional information alone does not suffice to explain functional load effects, there must be other mechanisms through which these effects operate. We hypothesize that one such mechanism could be perception.

Boomershine and colleagues (2008) found in a series of experiments that with the same stimuli, English listeners rate the sounds [d] and [ɾ] as more similar to one another than [ð], while Spanish listeners rate [d] and [ð] as more similar to one another than [ɾ]. They propose that this reflects the phonemic status of the three sounds in these two languages. [d] and [ɾ] are both allophones of /d/ in English, but /ð/ is a separate phoneme. In Spanish, it is [d] and [ð] that are allophones of /d/, while /ɾ/ is a separate phoneme. Allophones in participants' native language are thus perceived as more perceptually similar than separate phonemes.

Phonemicity is one level of contrastiveness, but it could be argued that functional load provides another, gradient, measure of contrastiveness among phonemes of a language. If that is so, higher functional load contrasts would be more perceptually distinct than lower functional load contrasts. To test this hypothesis, we examine in the following section whether perceptual confusability correlates with functional load measures in English and French.

1. Methods

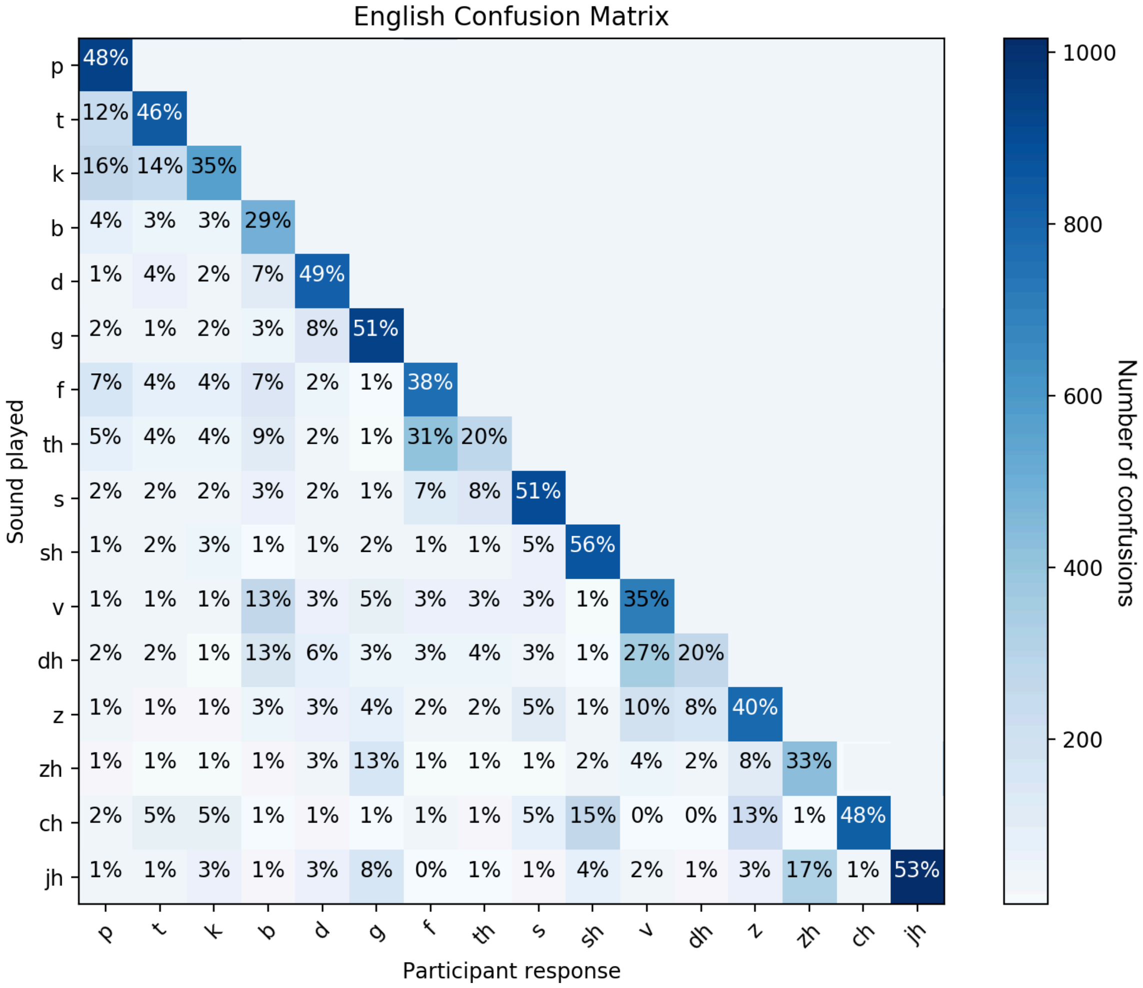
In order to obtain a measure of perceptual confusability of English sounds, we used data from perceptual confusions in noise for 14 English consonants (Wang & Bilger 1973). In their experimental task, native English listeners heard a consonant sound in a variety of positions in a syllable, and chose among all possible consonant options which consonant they had heard. Figure 10 adapted from Wang and Bilger (1973) using Scikit-learn (Pedregosa et al. 2011) plots the matrix for 14 English consonant sounds. In each row, the matrix shows the number of times a given sound is misperceived as another (in the intersecting column). This means that for any pair of sounds x and y, there are actually two values of confusion: the number of times participants misheard x as y, and the number of times they misheard y as x.



	p	t	k	b	d	g	f	th	s	sh	v	dh	z	zh	ch	jh
p	55%	12%	11%	1%	1%	3%	6%	2%	2%	1%	1%	1%	2%	1%	1%	1%
t	14%	49%	12%	2%	4%	2%	4%	2%	3%	1%	1%	1%	1%	1%	2%	1%
k	19%	15%	33%	2%	2%	3%	5%	2%	2%	2%	1%	0%	1%	0%	6%	6%
b	8%	4%	4%	29%	4%	1%	11%	7%	4%	0%	19%	6%	2%	0%	0%	1%
d	2%	3%	2%	10%	48%	8%	3%	2%	3%	1%	6%	5%	4%	1%	1%	2%
g	1%	1%	1%	5%	9%	55%	1%	1%	1%	1%	8%	2%	3%	2%	0%	9%
f	12%	5%	4%	5%	2%	1%	45%	10%	7%	1%	3%	2%	2%	0%	0%	1%
th	6%	4%	4%	6%	1%	1%	39%	16%	12%	2%	3%	2%	2%	0%	1%	1%
s	3%	2%	2%	3%	2%	1%	9%	4%	53%	3%	4%	2%	9%	1%	1%	1%
sh	2%	2%	3%	1%	2%	2%	2%	1%	5%	51%	1%	0%	2%	2%	17%	8%
v	2%	1%	2%	12%	3%	3%	5%	4%	3%	1%	41%	11%	9%	1%	0%	2%
dh	2%	2%	1%	14%	5%	4%	3%	4%	3%	1%	31%	16%	11%	1%	1%	2%
z	1%	1%	1%	4%	3%	7%	2%	2%	2%	1%	13%	7%	46%	4%	0%	5%
zh	1%	1%	1%	2%	4%	17%	1%	1%	1%	1%	5%	2%	8%	25%	0%	30%
ch	4%	9%	5%	1%	1%	1%	3%	1%	8%	13%	1%	0%	1%	1%	48%	2%
jh	1%	1%	1%	2%	4%	9%	0%	0%	1%	2%	2%	0%	6%	9%	1%	60%

***Figure 10**. Confusion matrix for 14 consonant sounds in English, adapted from Wang & Bilger (1973)*

Meanwhile, for each pair of sounds, there is only one value of functional load in each of the two measures we are comparing. To correlate this value with the confusability of the sounds involved, we then used the average of the two confusion values for each sound pair. This gave us the halved matrix in Figure 11 showing single confusion values for each sound pair.



	p	t	k	b	d	g	f	th	s	sh	v	dh	z	zh	ch	jh
p	48%															
t	12%	46%														
k	16%	14%	35%													
b	4%	3%	3%	29%												
d	1%	4%	2%	7%	49%											
g	2%	1%	2%	3%	8%	51%										
f	7%	4%	4%	7%	2%	1%	38%									
th	5%	4%	4%	9%	2%	1%	31%	20%								
s	2%	2%	2%	3%	2%	1%	7%	8%	51%							
sh	1%	2%	3%	1%	1%	2%	1%	1%	5%	56%						
v	1%	1%	1%	13%	3%	5%	3%	3%	3%	1%	35%					
dh	2%	2%	1%	13%	6%	3%	3%	4%	3%	1%	27%	20%				
z	1%	1%	1%	3%	3%	4%	2%	2%	5%	1%	10%	8%	40%			
zh	1%	1%	1%	1%	3%	13%	1%	1%	1%	2%	4%	2%	8%	33%		
ch	2%	5%	5%	1%	1%	1%	1%	1%	5%	15%	0%	0%	13%	1%	48%	
jh	1%	1%	3%	1%	3%	8%	0%	1%	1%	4%	2%	1%	3%	17%	1%	53%

Figure 11*. Halved confusion matrix for 14 consonant sounds in English, adapted from Wang & Bilger (1973). For any pair of sounds (e.g. /p/ and /t/), we took the average of how often one was heard as the other and vice versa to represent the average confusability of the two sounds (e.g. the average of how often /p/ was heard as /t/, and how often /t/ was heard as /p/)*

Values for English functional load were computed in the previously reported corpus study, for both entropy and minimal pair measures. Since neither measure was normally distributed, we used once again Spearman's correlation to test for a relationship between functional load and confusability.

Even if we do find that perceptual confusability and functional load are correlated in English, this could be merely an effect of acoustic distinctiveness. It could be that sound pairs that are acoustically more distinct to the human ear (or the mammalian auditory system, for that matter) are used more widely to establish lexical contrasts. If that were the case, this effect would not be language-specific, and we should find that functional load in any language could correlate with the perceptual confusability of English listeners. To check whether this is the case, we used the previously computed French functional load values as a control: if functional load does indeed influence perceptual confusability, English but not French functional load should correlate with English confusability.

2. Results

a. English functional load, English confusability

We found that both the entropy and the minimal pair measure of English functional load correlated significantly with the average number of confusions of English phonemes. However, the correlations ran in opposite directions for the two measures. While minimal pairs count correlated negatively with the number of confusions ($r_{s\ minpairs} = -0.44$, $p_{minpairs} < 0.01$), meaning that higher functional load contrasts according to minimal pairs were less confusable, entropy correlated positively with the number of confusions (n = 120, $r_{s\ entropy} = 0.24$, $p_{entropy} = 0.009$). Figure 12 shows these results.

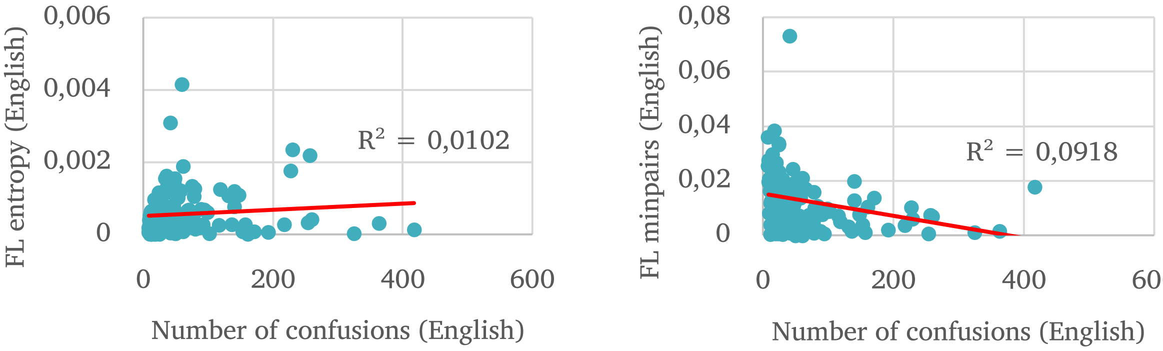


Figure 12*. Correlations between functional load measures in English and average number of confusions in English according to entropy (left) and minimal pairs (right)*

While both correlations are highly significant, their opposite directions cast doubts on the nature of the relationship between functional load and confusability. If both correlations had been negative, this would support our initial hypothesis that higher functional load contrasts are perceptually more distinct. We know from the previous tests on the two measures of functional load in English that the two measures, while correlated, exhibit different distributions. This difference appears once more in this latest result.

As these correlations are highly significant, it seems unlikely that they arose by chance. But could it be that they can merely be attributed to universal acoustic effects? To answer this question, we turn to our control case: does French functional load correlated with English confusability in the same way?

b. Control: French functional load, English confusability

In contrast, we found that neither the entropy nor the minimal pair measure of French functional load correlated significantly with the average number of confusions of English phonemes, for those sound contrasts that exist in both languages ($n = 66$, $r_{s\,entropy} = 0.04$, $p_{entropy} = 0.76$; $r_{s\,minpairs} = 0.11$, $p_{minpairs} = 0.37$). Figure 13 shows these results.

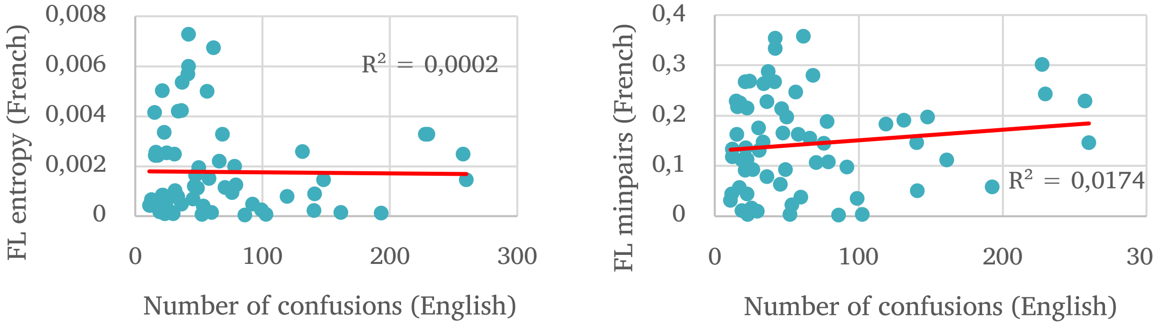


***Figure 13**. Correlations between functional load measures in French and average number of confusions in English according to entropy (left) and minimal pairs (right)*

This suggests that the correlations we find between English functional load and the average number of confusions in English are indeed language-specific, and not attributable to a universal preference to use acoustically distinct sounds to establish lexical contrasts.

3. Chapter conclusion

In the previous chapter, we found that the two measures of functional load are well correlated in real language corpora, but exhibit significant differences in how they rank sound contrasts within a language. In this chapter, we tested for a correlation between perception and functional load as a metric of lexical contrastiveness. While we find evidence for a significant correlation between perceptual confusability and functional load according to either measure within a language (English), the correlation runs in opposite directions for the two measures. This does not straightforwardly support our hypothesis that high functional load contrasts are easier to tell apart, but our control case

(French) shows that these correlations cannot be attributed merely to a universal preference for acoustic distinctiveness in lexical contrasts.

Another possibility for this ambiguous result stems from the nature of the task used to establish the confusion matrix we relied on. For each sound they heard, participants were given the full range of consonant options, making this a very open-ended task. Each cell in the full confusion matrix means "the number of times the sound x was heard as y among all 14 other possible consonants". In real language processing, on the other hand, when a listener hears a word with one ambiguous segment X in English, for instance [maʊX], the lexicon constrains the possibilities for X to [s] or [θ]. Since functional load is a lexical measure, such an open-ended task might not be the best way to assess its relationship to perceptual distinctiveness. To examine further the relationship between functional load, contrastiveness, and perceptual distinctiveness, we conduct a perceptual experiment and an artificial grammar learning experiment using a set of English sound contrasts for which the two measures of functional load make different predictions. We will see that a more restrictive perceptual task provides evidence that this apparent contradiction might come from the different rank ordering of contrasts according to functional load in the two measures, and, furthermore, that phonological learning in an artificial grammar learning experiment impacts perceptual distinctiveness depending on differences in functional load.

Chapter IV. Perceptual discrimination and artificial grammar learning experiments

In the previous chapter, we speculated that one mechanism that could account for a better diachronic preservation of high-functional-load contrasts could be the perceived similarity of sounds in L1. However, existing data on perceptual similarity did not provide a straightforward answer, as correlations with functional load, while highly significant, did not follow a consistent direction in the two measures of functional load. This could have been due to differences between the two measures, but also to the open-ended nature of the task used to obtain the previous measures of confusability.

In order to obtain a measure of confusability that would only quantify how often two sounds X and Y are mistaken for each other by English listeners, we first conducted a perceptual discrimination experiment using an AX task. When hearing in noise a minimal pair of non-words hinging on a given contrast in English (e.g. [ata] – [aʧa]), how often do native speakers confuse the two?

In parallel, we also wanted to examine the effect of functional load on learning phonological alternations. Previous studies in laboratory phonology experiments have shown that participants are fairly successful at acquiring phonological alternation patterns in artificial languages (Moreton & Pater, 2012a, 2012b for an overview). However, recent work by Yin and White (2018) showed that neutralizing alternations are more difficult to learn than non-neutralizing ones if they induce homophones in the artificial language. This suggests that not all alternations are equally learnable.

If an alternation in the artificial language involves sounds with a high-functional-load contrast in the learner's native language, would it be easier or more difficult to learn? On the one hand, it is important for a listener to tell apart sounds, even more so for a high-functional-load contrast, and this in turn could make the alternation easier to learn. On the other hand, if these sounds are so distinct, mapping them onto a single phoneme might be more difficult.

Not all types of alternations induce the same amount of change in terms of information. Say, for instance, that a speaker has two phonemes /X/ and /Y/ realized respectively as [X] and [Y] in their native language. They are learning an L2 in which these sounds are involved in an alternation. It could be that [X] and [Y] are both allophones of /X/ in L2 (and there is no phoneme /Y/ in L2), or it could be that /X/ and /Y/ are realized as [X] and [Y] in some contexts, but in some other ones they neutralize to [X] (Figure 14). In the first case, they would be learning a non-neutralizing alternation. Since there was never a phoneme /Y/ in L2, there is no loss of information within the L2. In the other case, a neutralizing alternation, /X/ and /Y/ are phonemes in L2, just as they were in L1, but realize as the same phone [X] under specific conditions. There is now some loss of information within L2 – some surface [X] might need to be reconstructed as underlying /Y/. Learning a non-neutralizing alternation means learning that the X-Y contrast from L1 does not matter in L2, but within L2, there is no information lost from realizing /X/ as [Y] in some contexts, as there is a single mapping from [Y] to /X/. However, learning a neutralizing alternation means learning that the X-Y contrast still matters in L2, but information is lost in contexts where it is neutralized. It also means then learning to reduce the functional load of the /X/-/Y/ contrast L2 compared to the L1 reference.

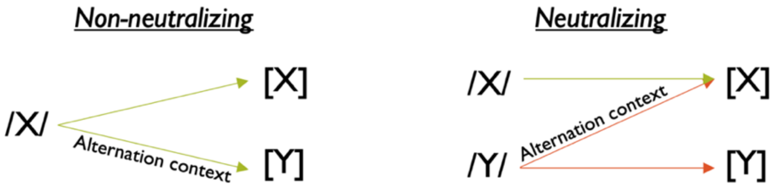


***Figure 14**. Phoneme to phones mapping in non-neutralizing and neutralizing alternations*

In Yin and White's (2018) analysis, all alternations of the same type were grouped together (neutralizing vs. non-neutralizing), and potential differences between alternations of the same type involving different sounds were not examined. Based on their results, if we do not induce homophones in an artificial language, we should not observe differences in learning regardless of whether sounds neutralize or not.

If functional load matters when learning alternations in another language, we should find that alternations involving different sound pairs are not learned to the same extent. Furthermore, if phonological status (phonemic or not) impacts the perceptual distinctiveness of sounds, we should observe a difference between a perceptual test before the artificial grammar learning task, and another perceptual test after.

1. Experiment 1, part 1: AX perceptual task, baseline test

a. Materials and design

In the perceptual task, participants heard two sounds (VC_1 – VC_2, or $V_1C_1V_2$ – $V_1C_2V_2$), and were asked to decide whether C_1 and C_2 were the same consonant. The two tokens in each pair were spoken by two different speakers, one female, one male, and were overlaid with multitalker babble (Wilson et al. 2007) at a signal to noise ratio of -10dB. The order in which the tokens were played was randomized for each participant, though the same tokens were paired (e.g. $[ata]_1$ – $[ata]_2$ vs. $[ata]_2$ – $[ata]_1$). The consonants tested were the following pairs: t-ʧ, s-ʃ, z-ʒ, b-f. Vowels were a, i, or u. Participants heard 12 random VC_1 – VC_2 and 36 $V_1C_1V_2$ – $V_1C_2V_2$ trials involving each pair of sounds combined with each vowel or pair of vowels, approximately half same, half different (total 192 trials, lasting about 15 minutes). A full list of all stimuli can be found in Appendix II. The functional load of the contrasts we examined were as follows (Figure 15).

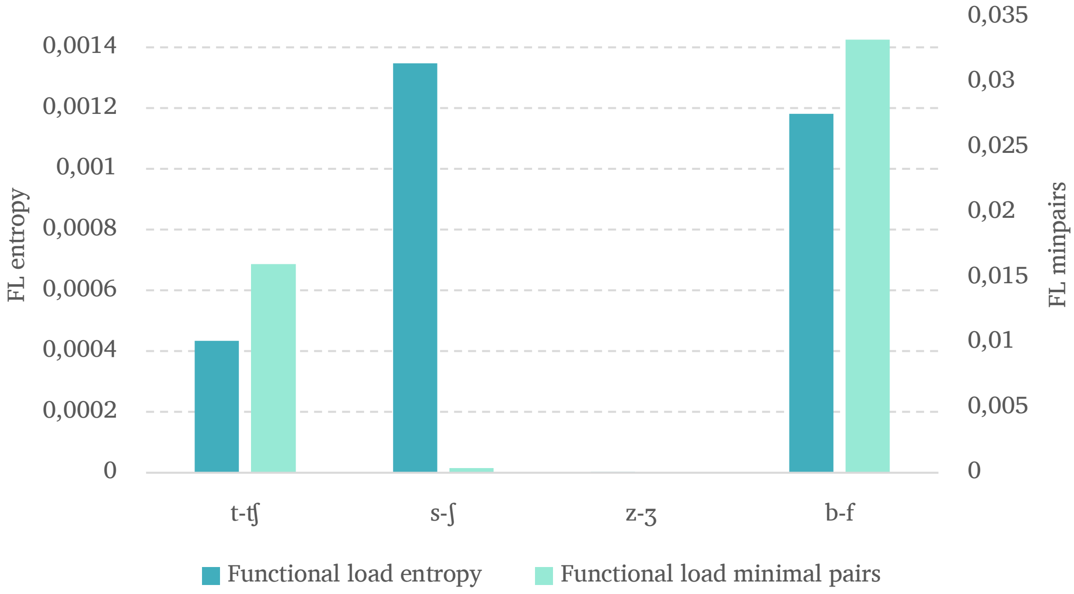


Figure 15. *Functional load of contrasts in the perceptual task*

These contrasts were chosen because the two measures of functional load ranked them differently. Another reason for the choice of these particular contrasts related to a second experiment, and will be explained in a later section. According to entropy, z-ʒ has the lowest functional load, followed by t-ʧ, b-f, then s-ʃ. According to minimal pair count, z-ʒ also has the lowest functional load, followed by s-ʃ, t-ʧ, then b-f. Figure 16 represents this difference in order.

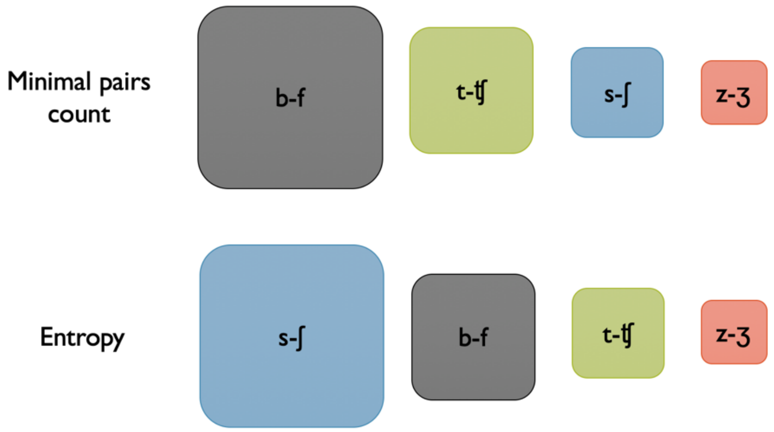


Figure 16. *Functional load rank ordering of test contrasts in the two measures*

This difference in rank ordering makes diverging predictions for the outcome of the perceptual task. If a given contrast's perceptual distinctiveness depends on its functional load, the two measures predict that participants should find different contrasts more perceptually distinct.

b. Procedure

Participants sat in front of a computer in a soundproof room, wearing noise-cancelling headphones. The experiment was conducted online using the Experigen experiment building platform (Becker & Levine 2010). Participants pressed the enter key on a keyboard to play each of the two tokens in an AX pair (e.g. [ata] by speaker 1, [ata] by speaker 2). They clicked buttons on screen with the mouse cursor to answer "Yes" or "No" to the question: "Was it the same consonant in both words?" (Figure 17). Each token could only be played once.

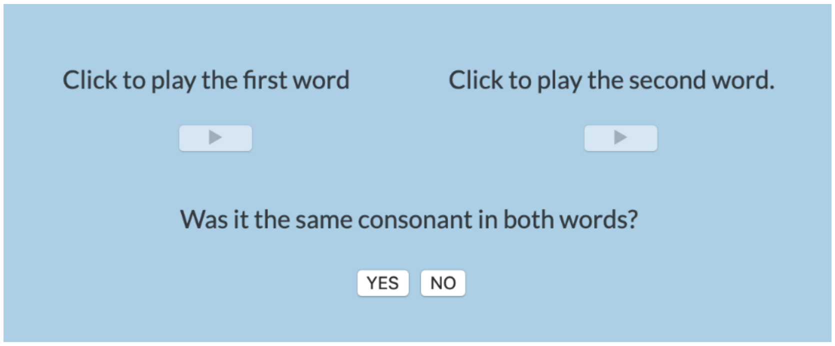


Figure 17*. Participant view during the perceptual task*

c. Participants

77 undergraduate students at UCLA participated in the baseline AX task (mean age = 21.62, SD = 3.95). 52 participants received course credit for their participation and 25 received 10USD in cash. Participants who received course credit were recruited through the UCLA Psychology Department Subject Pool, and paid participants through flyers on the UCLA campus.

d. Data analysis

Data was analyzed using linear mixed effect models with the nlme package in R (R Core Team 2018). The dependent variable was participants' d′ sensitivity index for each of the sound pairs[10]. The only fixed effect was Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f). Random effect structure was kept maximal under condition of convergence (random intercepts and slopes for Participants).

2. Results

Two participants were excluded for reporting after the study that they had experienced speech or hearing impairments (2.6% of the data).

We examined whether participants were more or less accurate in distinguishing the members in each of the tested sound pairs (t-ʧ, s-ʃ, z-ʒ, b-f). Figure 18 shows participants' d' for each of the sound pairs. The higher the d', the more participants were perceptually sensitive to the difference between two members of a sound pair. Additionally, d' measures take into account response biases (in case participants preferred clicking one response button over the other)

---

[10] d' measures were computed using the following definitions:
- hit: tokens had different consonants (e.g. [t] and [ʧ]) and participant correctly said they were different
- false alarm: tokens had different consonants (e.g. [t] and [ʧ]) and participant incorrectly said they were the same
- correct rejection: tokens had the same consonant (e.g. [t] and [t]) and participant correctly said they were the same
- miss: two tokens had the same consonant (e.g. [t] and [t]) and participant incorrectly said they were different

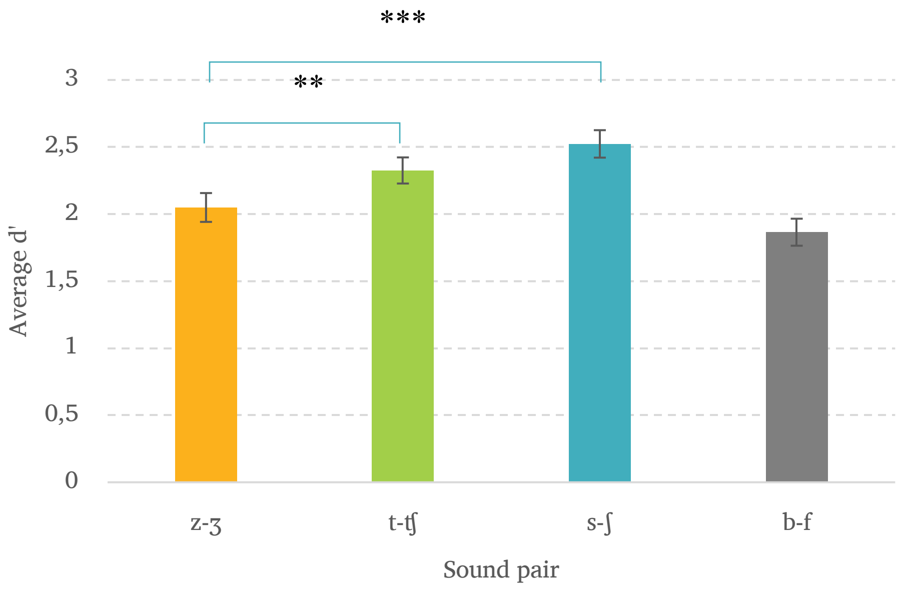


Figure 18. *d' index depending on sound pair, English listeners baseline perception*

Overall, z and ʒ were more difficult to tell apart than t-ʧ and s-ʃ, but comparable to b-f. Table 4 summarizes the model coefficients.

Table 4. *Coefficients in a model fitted to d' in the baseline perceptual task, fixed effect of sound pair (random intercepts and slopes for Participants)*

		Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
Intercept (b-f)		1.93	0.11	219	17.47	<0.001	***
Sound pair	s-ʃ	0.59	0.10	219	5.85	<0.001	***
	t-ʧ	0.39	0.10	219	4.10	<0.001	***
	z-ʒ	0.12	0.10	219	1.18	0.23	

Pairwise comparisons using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018)) confirmed that b-f was not significantly different from z-ʒ. t-ʧ and s-ʃ both had higher d' values than b-f and z-ʒ, but t-ʧ and s-ʃ were only marginally different from one another. This is summarized in Table 5.

Table 5. *Coefficients for pairwise comparisons of sound pairs in the baseline perceptual task*

	Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
b-f / z-ʒ	-0.12	0.10	219	-1.18	0.64	
s-ʃ / z-ʒ	0.47	0.09	219	5.30	<0.0001	***
t-ʧ / z-ʒ	0.28	0.08	219	3.33	0.006	**
s-ʃ / t-ʧ	0.20	0.08	219	2.50	0.06	.

These results can be schematized as:

s-ʃ (>?) **t-ʧ** > **z-ʒ, b-f**

where > means "more perceptually distinct than". Setting aside for now the case of b-f, and if we conservatively disregard the marginal difference between t-ʧ and s-ʃ, this order is actually consistent with the functional load ranking according to both entropy and minimal pair count (in Figure 16 repeated below). The contrast with the lowest functional load in both measures, z-ʒ, is one that English listeners find harder to perceive accurately. If the difference between s-ʃ and t-ʧ were actually significant, the order would be more consistent with the entropy measure.

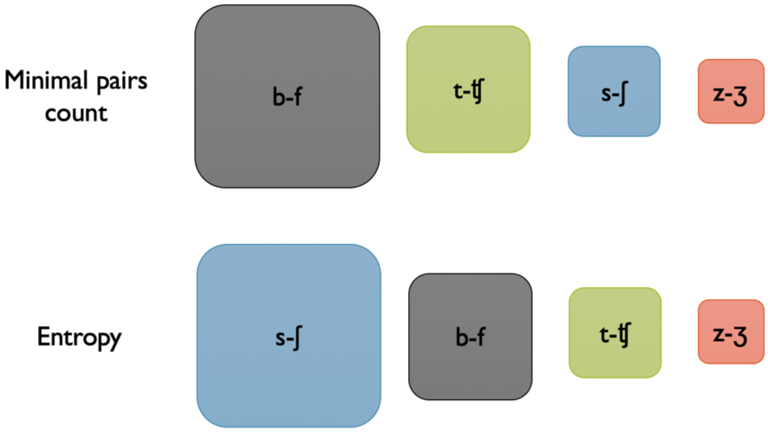


Figure 16 (repeated). *Functional load rank ordering of test contrasts in the two measures*

In terms of functional load, the b-f contrast is quite important according to both measures. What could explain then why b and f were as hard to tell apart as the lowest functional load z-ʒ contrast? This result is in fact consistent with Wang and Bilger's (1973) confusion matrix data, used in the previous chapter to correlate functional load to perceptual distinctiveness: on average, z-ʒ were mistaken one for the other in 8% of the cases, and b-f in 7% (over ¾ of sound pairs from the confusion matrix had less than 6% confusability). Of course, aside from functional load, acoustic similarity must play a large role – the acoustic similarity of b-f could be so great that this could not be overridden by the importance of the contrast in terms of functional load. But if acoustic similarity accounts for d' results of the b-f distinction, it could also explain the outcome for the three other contrasts, or at least we cannot disentangle the effect of acoustic similarity from that of functional load. In order to achieve this, we will need to rely on the results of the artificial grammar learning task and the second perceptual task.

3. Interim conclusion

The previously described perceptual discrimination task was designed to establish a baseline of the discriminability of four pairs of segmental contrasts for English listeners. We found that the lowest functional load contrast according to both measures, z-ʒ, also had the lowest perceptual discriminability. There was only a marginally significant difference between two of the higher functional load contrasts, s-ʃ and t-ʧ, and thus no clear-cut evidence in support of either of the two measures against the other. Additionally, b-f should have a relatively high functional load, but is as difficult to tell apart as z and ʒ. This is consistent with confusion matrix data from previous studies.

In their study on the perceptual similarity of fricatives in Dutch and English, Johnson and Babel (2008) concluded that phonetic similarity is comprised of three components: auditory similarity, the language's phonetic inventory, and language-specific patterns of alternation. The latter two categories both contribute to functional load. We could conjecture that despite their relatively high functional load in English, b and f are just acoustically too similar, and this explains why they were not significantly different from the lowest functional load z-ʒ in the baseline perceptual task. This forces us however to consider that the differences we found among the other sound pairs might also be due to pure acoustic (dis/)similarity.

Examining corpus data and baseline perceptual discrimination for English listeners allowed us to test hypotheses about functional load and its role in the current, stable state of the language. To examine whether contrastiveness in L1 affects the learning of phonological patterns in other languages, we also conducted an artificial grammar learning task. Would alternations involving high functional load sound pairs in L1 be easier or more difficult to learn in an artificial language? Does it matter whether these alternations are neutralizing or not?

The artificial grammar learning experiment (henceforth AGL) would also allow us to test the effect of learning an alternation on perceptual distinctiveness. If the distinctiveness of our test sound pairs does not change uniformly after learning an artificial language, this would be an effect of learning the alternations from the AGL, and would provide evidence towards a role of functional load in phonological learning and/or perceptual distinctiveness. Additionally, changes in perceptual distinctiveness following the AGL cannot be attributed to pure acoustic (dis/)similarity, and would then have to come from effects of phonological learning and/or functional load.

4. Experiment 2: Artificial grammar learning experiment

In the artificial grammar learning experiment, three groups of participants were each exposed to one of three artificial languages. Each language had two neutralizing alternations and one non-neutralizing alternation. We built our experiment on previously used methodology (White, 2014, Yin & White, 2018). The alternations in our study were tested in Yin and White's experiments (z-ʒ, t-ʧ and s-ʃ), though the neutralizing alternations had not been compared among themselves in their study. They had found that non-neutralizing alternations were overall better learned than neutralizing ones only if neutralizing alternations induced homophony.

In our study, we tried to examine whether this holds regardless of the functional load of the neutralized contrast in participants' native language. If speakers are sensitive to differences in functional load in their L1 and tend to preserve high-functional-load contrasts, we expected that they would learn alternations involving a low-functional-load contrast better than those involving a higher FL contrast.

a. Materials and design

Participants were exposed to one of three artificial languages. We examined three pairs of sounds, t-ʧ, s-ʃ and z-ʒ. These pairs were chosen because they were pre-existing phonemes in English, their contrasts carried different functional loads, and they are crosslinguistically involved in the same type of common alternation (palatalization, identified in over 100 languages from various families, Bateman, 2007).

The functional load of these three sound pairs in English is repeated below (Figure 15, repeated), with the addition of the sound pair b-f. This pair was tested in the baseline perceptual task, but are not involved in an alternation in the artificial grammar learning task, though the pair appeared in non-alternating fillers.

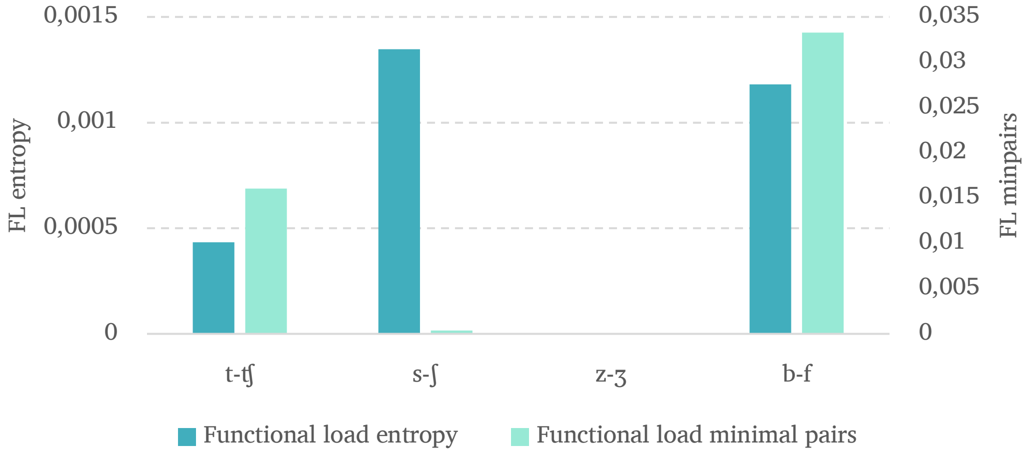


Figure 15 *(repeated). Functional load of contrasts in perceptual task*

The three artificial languages were constructed so that each language had a non-neutralizing alternation involving one sound pair (e.g. /s/ → [ʃ]/_i, where [ʃ] does not occur anywhere else in the data: no evidence of /ʃ/ as a phoneme in the language) and two neutralizing alternations involving the two other sound pairs (e.g. /t/→[ʧ]/_i and /z/→ [ʒ]/_i , where /ʧ/ and /ʒ/ are also phonemes in the language). We will henceforth refer to the three language conditions using the phoneme that does not neutralize in each language (non-neutralizing /t/, non-neutralizing /s/ or non-neutralizing /z/). Taken together, these three language conditions allowed us to examine each sound pair in a neutralizing alternation and a non-neutralizing alternation (e.g. /s/ → [ʃ]/_i, where [ʃ] is not a phoneme, and when /ʃ/ is a phoneme).

Singular words in the artificial language were of the form CVCVC, and the plural form was created by adding a suffix /-i/. All words, singular and plural, bore penultimate stress. Words could end in an alternating test sound (t, z, or s), and had then a plural form in which the final consonant was palatalized (e.g. zuvot - zuvoʧi, dunis - duniʃi, sagoz - sagoʒi), or in a non-alternating test sound (ʧ, ʃ or ʒ) or filler sound (p, b, d, k, g, f or v). Non-alternating test sounds and filler sounds did not alternate in the plural forms (e.g. gavoʧ - gavoʧi, gubid – gubidi).

To make all artificial languages comparable, the functional load of each test sound pair within the artificial language was equal, and null. This is because even when a test pair like /s/-/ʃ/ was contrastive in the artificial language (neutralizing to [ʃ] in front of /i/), the language was constructed so that no minimal pairs hinged on /s/-/ʃ/. /s/ and /ʃ/ appeared in the same environments, except before /i/. This also prevented the creation of homophones by neutralizing the contrast. Since minimal pairs are necessary for either functional load measure to be non-null, all tested contrasts had a null functional load in the artificial language. What mattered, then, if we observed a difference between sound pairs, was that each pair had a given functional load in English, the participants' native language, and that the alternations taught in the AGL involved different changes to how much information was carried by each contrast.

In the exposure phase, participants heard 8 pairs of words ending in each of the test sounds, as well as 5 pairs of words ending in each of the fillers sounds (total 72 pairs). They heard each word pair twice.

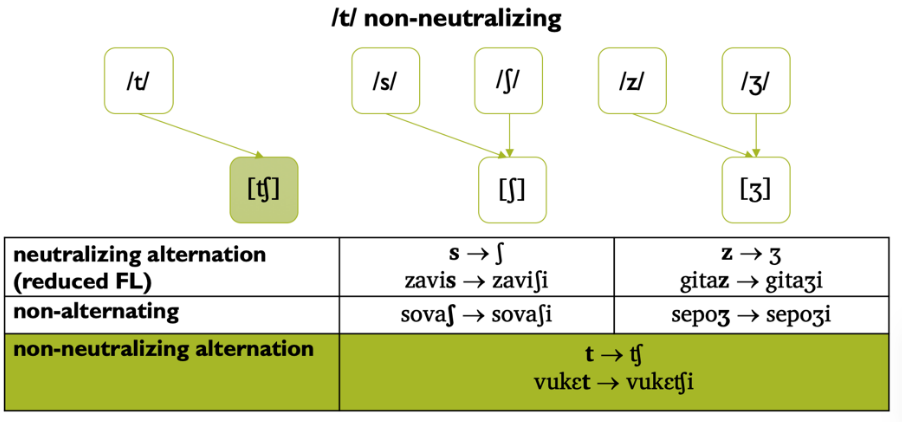
In the verification phase, participants were asked to choose the correct plural form for 25 test words heard in the exposure phase, 5 ending in each of the test sounds, and 5 filler words randomly chosen among fillers from the exposure phase.

In the generalization phase, participants were asked to choose the correct plural form for 50 novel test words, 10 ending in each of the test sounds and 10 filler words randomly chosen among 50 possible filler words ending in one of the fillers sounds. Correct and incorrect plural options followed the same pattern as the verification phase.

All stimuli were phonotactically well-formed in English. A full list of all stimuli can be found in Appendix II.

- *Non-neutralizing /t/ language*

In the non-neutralizing /t/ language, /t, s, ʃ, z, ʒ/ were phonemes. Before [i], /t/ palatalized to its allophone [ʧ], but /s/ and /ʃ/ neutralized to [ʃ] and /z/ and /ʒ/ neutralized to [ʒ]. This pattern is summarized in Figure 19 below:



neutralizing alternation (reduced FL)	s → ʃ zavis → zaviʃi	z → ʒ gitaz → gitaʒi
non-alternating	sovaʃ → sovaʃi	sepoʒ → sepoʒi
non-neutralizing alternation	t → ʧ vukɛt → vukɛʧi	

Figure 19*. Sound patterns in the non-neutralizing /t/ language*

In the verification and generalization phases, plural options were given as follows in Table 6:

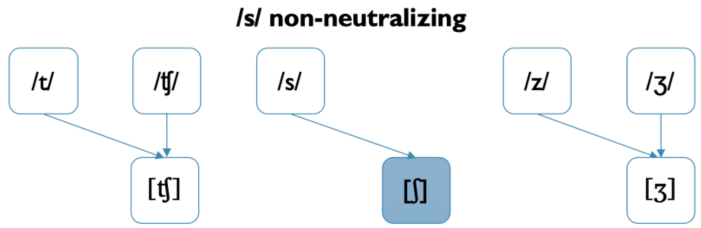
Table 6. *Correct and incorrect plural options in the non-neutralizing /t/ language. Filler sound options remain identical across languages.*

	Final consonant in singular form	Correct plural consonant	Incorrect plural consonant
Test sounds	ʃ	ʃ	s
	ʒ	ʒ	z
	s	ʃ	s
	z	ʒ	z
	t	tʃ	t
Fillers sounds	p	p	tʃ
	b	b	ʒ
	d	d	ʒ
	k	k	tʃ
	g	g	ʒ
	f	f	ʃ
	v	v	ʃ

This language provided data for the following conditions: non-neutralizing t-tʃ, neutralizing s-ʃ and neutralizing z-ʒ.

- *Non-neutralizing /s/ language*

In the non-neutralizing /s/ language, /t, ʧ, s, z, ʒ/ were phonemes. Before [i], /s/ palatalized to its allophone [ʃ], but /t/ and /ʧ/ neutralized to [ʧ] and /z/ and /ʒ/ neutralized to [ʒ]. This pattern is summarized in Figure 20 below:



neutralizing alternation (reduced FL)	t → ʧ zavit → zaviʧi	z → ʒ gitaz → gitaʒi
non-alternating	sovaʧ → sovaʧi	sepoʒ → sepoʒi
non-neutralizing alternation	s → ʃ vukɛs → vukɛʃi	

Figure 20. *Sound patterns in the non-neutralizing /s/ language*

In the verification and generalization phases, plural options were given as follows in Table 7.

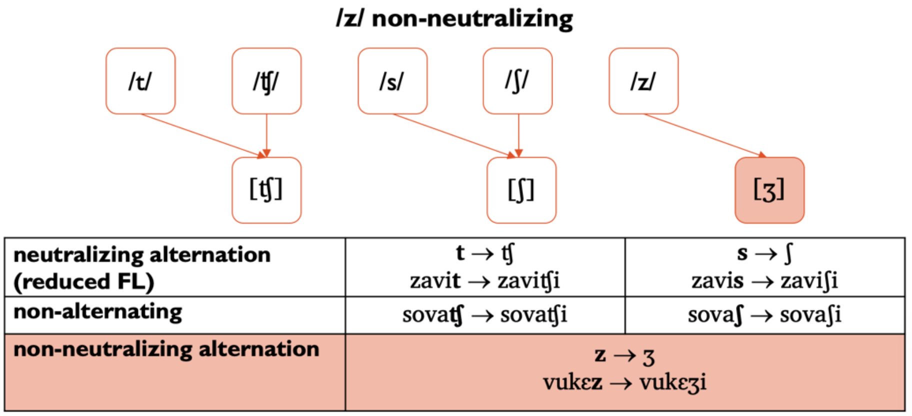
Table 7*. Correct and incorrect plural options in the non-neutralizing /s/ language. Filler sound options remain identical across conditions.*

	Final consonant in singular form	Correct plural consonant	Incorrect plural consonant
Test sounds	tʃ	tʃ	t
	ʒ	ʒ	z
	s	ʃ	s
	z	ʒ	z
	t	tʃ	t
Fillers	p	p	tʃ
	b	b	ʒ
	d	d	ʒ
	k	k	tʃ
	g	g	ʒ
	f	f	ʃ
	v	v	ʃ

This language provided data for the following conditions: non-neutralizing s-ʃ, neutralizing t-tʃ and neutralizing z-ʒ.

- *Non-neutralizing /z/ language*

In the non-neutralizing /z/ language, /t, ʧ, s, ʃ, z/ were phonemes. Before [i], /z/ palatalized to its allophone [ʒ], but /t/ and /ʧ/ neutralized to [ʧ] and /s/ and /ʃ/ neutralized to [ʃ]. This pattern is summarized in Figure 21 below:



neutralizing alternation (reduced FL)	**t → ʧ** **zavit → zaviʧi**	**s → ʃ** **zavis → zaviʃi**
non-alternating	**sovaʧ → sovaʧi**	**sovaʃ → sovaʃi**
non-neutralizing alternation	**z → ʒ** **vukɛz → vukɛʒi**	

Figure 21*. Sound patterns in the non-neutralizing /z/ language*

In the verification and generalization phases, plural options were given as follows in Table 8.

Table 8. *Correct and incorrect plural options in the non-neutralizing /z/ condition. Filler sound options remain identical across conditions.*

	Final consonant in singular form	Correct plural consonant	Incorrect plural consonant
Test sounds	tʃ	tʃ	t
	ʃ	ʃ	s
	s	ʃ	s
	z	ʒ	z
	t	tʃ	t
Fillers	p	p	tʃ
	b	b	ʒ
	d	d	ʒ
	k	k	tʃ
	g	g	ʒ
	f	f	ʃ
	v	v	ʃ

This language provided data for the following conditions: non-neutralizing z-ʒ, neutralizing t-tʃ and neutralizing s-ʃ.

b. Procedure

Participants sat in a soundproof room, in front of a desktop computer, and wore noise-cancelling headphones. The experiment was conducted online using the Experigen experiment building platform (Becker & Levine 2010). Participants were told that they would be hearing words from a foreign language: they did not need to memorize them, but were encouraged to repeat them, and would be asked opinions about these words and others later. During the exposure phase, participants pressed the enter key on a keyboard to hear a word while a picture of a singular or plural object appeared on the screen (e.g. [gavis] with a picture of a bird, or [gaviʃi] with a picture of two birds). All words in the exposure phase were presented in random order, so that the other word of the pair (e.g. the plural form) did not necessarily appear immediately before or after the first one (e.g. the singular form). This random order was chosen as a result of participants performing at ceiling level in pilot experiments where plural forms immediately followed the corresponding singulars. Figure 22 shows an example of a random sequence of training trials.

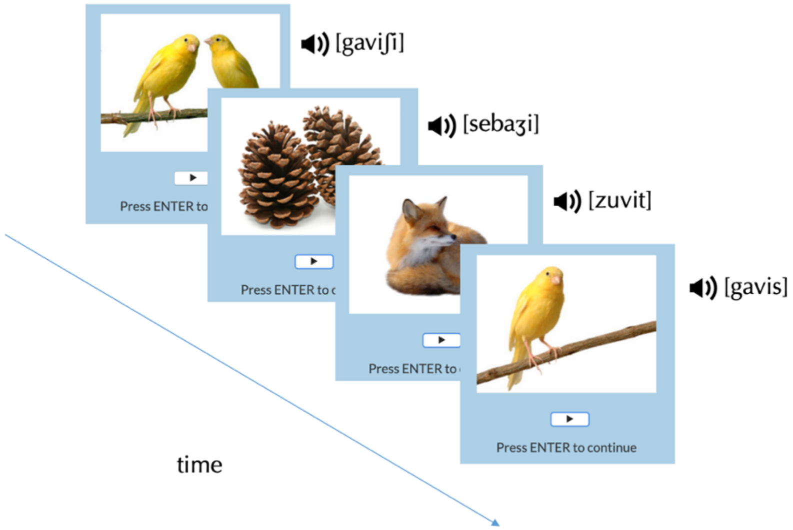
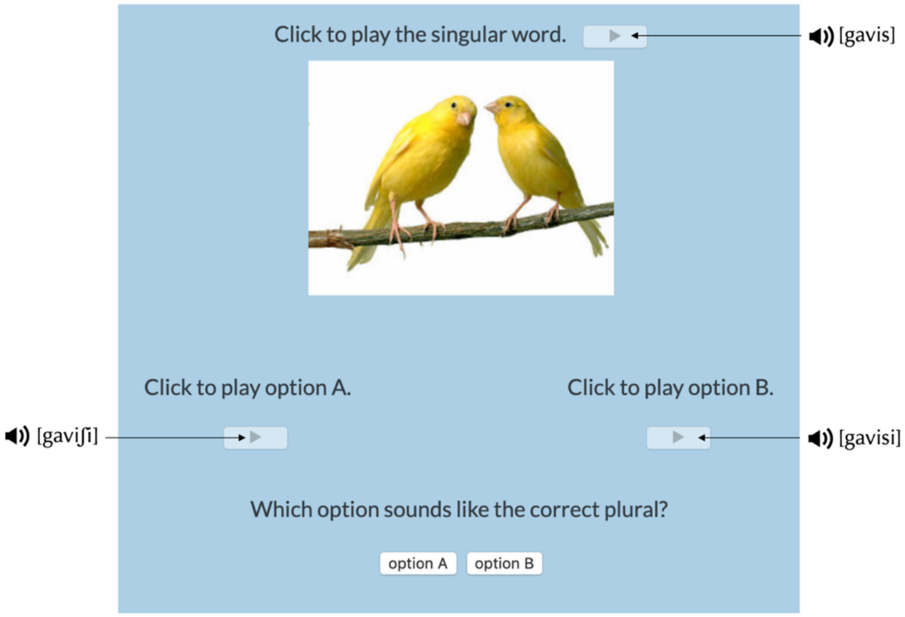


Figure 22. Structure of a training block

There were two blocks of training trials – each participant heard each training word pair (singular and plural) twice.

During the verification phase, participants completed a forced-choice task on a subset of the words they had heard during the exposure phase. Participants heard a singular word (e.g. [gavis]), then saw the plural picture and were asked to click buttons on screen to choose between two options for the plural word (e.g. *_[gavisi]_ or [gaviʃi]). Figure 23 shows the experimental screen for a verification trial. Participants clicked the on-screen buttons or pressed the enter key to play the singular form and the two plural alternatives, then clicked the on-screen buttons to indicate their choice for the correct plural form.



Click to play the singular word.

Click to play option A.

Click to play option B.

Which option sounds like the correct plural?

Figure 23. *Verification trial. In this trial, option A is the correct plural.*

After the verification phase, participants completed a generalization phase, during which they were tested on novel words. Generalization trials were similar to verification trials, except that no image was shown, as these words had not been previously heard and associated with images.

c. Participants

156 native speakers of American English (self-reported) participated in the experiment (mean age 21.95, SD = 5.39). Participants were undergraduate students at UCLA, and received either course credit (n = 106) or $10 for their participation (n = 50). None had any known history of speech of hearing impairment (self-reported prior to the experiment).

d. Data analysis

Data was analyzed using a mixed effects logistic regression with the lme4 package in R (R Core Team 2018). The dependent variable was whether participants chose the correct plural options for singular words. The fixed effects were Phase (verification, generalization), Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f), and Condition (neutralizing or non-neutralizing). Random effect structure was kept maximal under condition of convergence (random intercepts and slopes for Participants and random intercepts for Items).

e. Results

Four participants were excluded for reporting after the study that they had experienced speech or hearing impairments (2.6% of the data).

We first ran a model including Phase (verification, generalization), Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f), and Condition (neutralizing or non-neutralizing), and all their interactions. There was no effect of Phase (verification, generalization): participants did not perform differently on the words they were trained on during the exposure phase or the new words from the generalization phase. We compared this model to its subset model excluding the Phase predictor using the anova() function in R(R Core Team

2018), and found no significant difference ($\chi^2(6) = 3.89$, $p = 0.69$). The results we present in the following thus exclude the Phase factor.

We took higher correct response rates to reflect better learning of the alternations. Figure 24 shows the average percentage of correct answers participants gave in neutralizing and non-neutralizing alternations for each of the sound pairs.

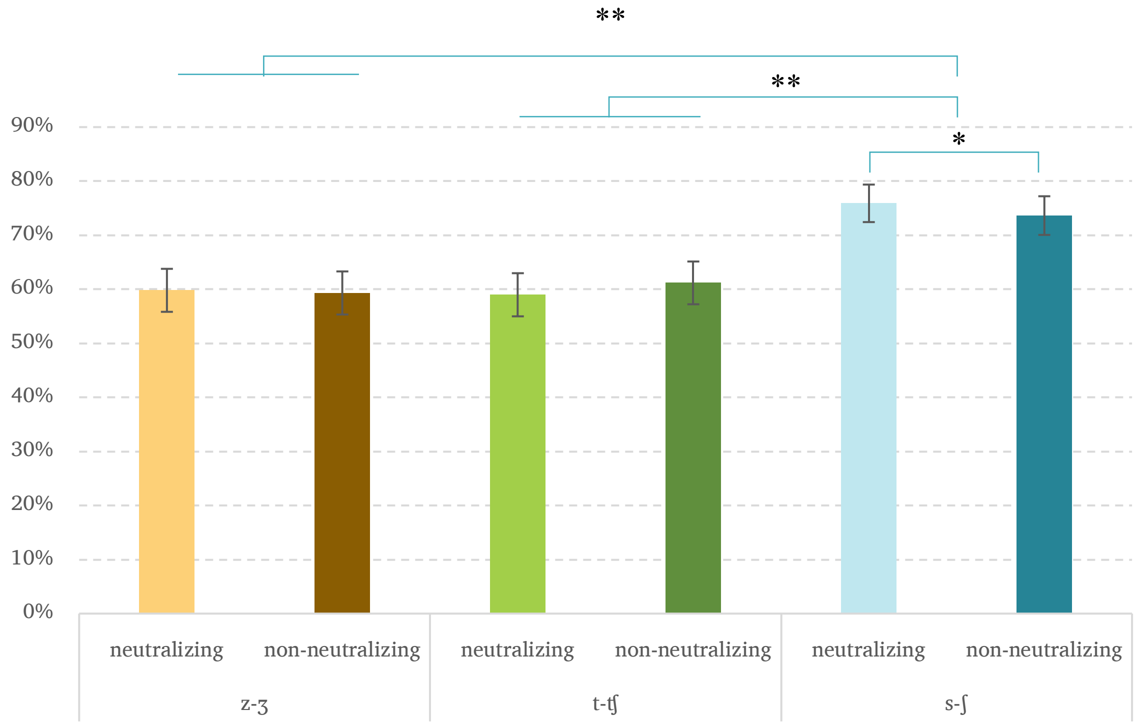


***Figure 24**. Average percentage of correct responses per sound pair per alternation in the artificial grammar learning task*

For t-ʧ and z-ʒ, we partially replicated Yin and White's result that in the absence of homophony, neutralizing alternations are learned to the same extent as non-neutralizing ones involving the same sound pairs. However, we also found a significant interaction between sound pair and alternation type: for the s-ʃ pair, a rather high-functional-load contrast according to both entropy and minimal pair count, the neutralizing alternation was learned better. This was confirmed in a pairwise comparison of means using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018)): (s-ʃ neutralizing vs. s-ʃ non-neutralizing, $p < 0.05$).

We also found a main effect of Sound Pair. Overall, alternations involving s-ʃ were learned the best, and alternations involving t-ʧ and z-ʒ were learned less well, with t-ʧ alternations only marginally better than z-ʒ alternations. Table 9 summarizes the coefficients and significance levels of the factors in our model.

***Table 9**. Coefficients in a logistic regression model fitted to correct/incorrect responses, fixed effects of alternation type, sound pair, and their interaction (random intercepts and slopes for Participants and random intercepts for Items)*

		Estimate	**Std. Error**	**z-value**	**p-value**	
Intercept (z-ʒ, neutralizing)		0.62	0.20	3.10	0.002	**
Alternation	non-neutralizing	0.01	0.13	0.10	0.92	
Sound pair	s-ʃ	0.84	0.26	3.23	0.001	**
	t-ʧ	-0.45	0.25	-1.79	0.07	
Interactions	Non-neutralizing : s-ʃ	-0.47	0.22	-2.18	0.03	*
	Non-neutralizing : t-ʧ	0.25	0.19	1.35	0.18	

These results can be schematized as:

s-ʃ > **z-ʒ** (>?) **t-ʧ**

where > means "alternations involving [sound pair on the left] were learned better than". The order of s-ʃ > t-ʧ is consistent with the functional load ranking based on entropy, but not minimal pair count (in Figure 16 repeated below).

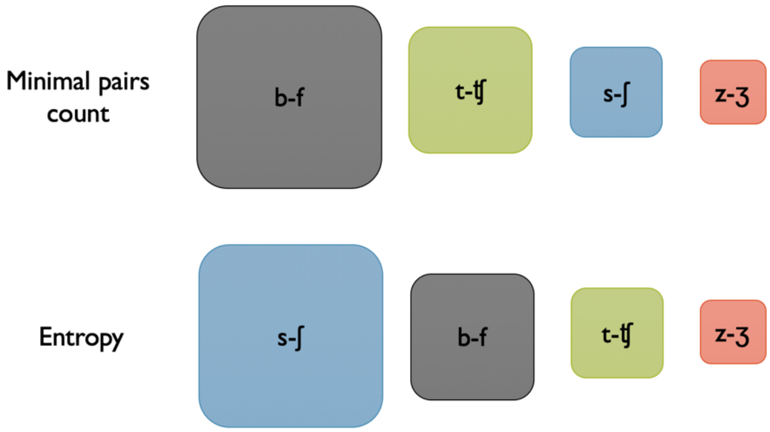


Figure 16 (repeated). *Functional load rank ordering of test contrasts in the two measures*

Recall that our baseline perceptual task yielded results that were consistent with both measures:

s-ʃ (>?) **t-ʧ** > **z-ʒ**

where > means "more perceptually distinct than". We had noted that if the difference between s-ʃ and t-ʧ in the baseline perceptual task was actually significant, the order would be more consistent with the entropy measure. We have now shown that difference through the artificial grammar learning task. The t-ʧ contrast patterns with the high functional load s-ʃ in the perceptual task, and with the low function load z-ʒ in the artificial grammar learning task, which would be consistent with a ranking in between the other two.

8. Interim summary

In the artificial grammar learning task, we tested whether alternations involving sound pairs with different functional loads were learned differently. All alternations involved palatalization triggered by a following [i]. All participants learned three alternations, two neutralizing and one non-neutralizing, with equal amounts of evidence for each alternation. Three artificial language conditions across participants allowed us to teach alternations involving each sound pair (t-ʧ, z-ʒ, s-ʃ) in a neutralizing and a non-neutralizing alternation. We partially replicated Yin and White's (2018) result that neutralizing alternations were not learned differently from non-neutralizing ones involving the same sound pairs: this was true for t-ʧ and z-ʒ, but not s-ʃ, where the neutralizing alternation was actually learned better. Though statistically significant, the difference in learning accuracy was small (2%) – more data would be needed to determine whether this effect is robust. Overall, the results from the AGL task rank the s-ʃ pair above t-ʧ and z-ʒ in terms of learning alternations. If we recall the baseline perceptual task ranked s-ʃ and t-ʧ above z-ʒ, we now have a tentative ranking of:

s-ʃ > **t-ʧ** > **z-ʒ**

which is consistent with the entropy measure but not the minimal pair count.

If contrastiveness affects perceptual distinctiveness, and learning phonological alternations involves learning new functional loads for native contrasts, would we observe a change in the perceptual distinctiveness of the sound pairs involved after the AGL task? We tested this using another AX task after the AGL.

9. Experiment 1, part 2: Post-AGL perceptual task

In our study, successfully learning to neutralize a contrast from L1 (e.g. t-ʧ) should entail temporarily reducing the functional load of the sounds involved: in the artificial language, this contrast is no longer active in some positions. If so, this might influence the perceptual similarity of the test sounds. We assess this by comparing the results of the pre-AGL perceptual task to a post-AGL perceptual task.

In a pilot version of the experiments, the artificial grammar learning task was both preceded and followed by a perceptual task involving the same sound pairs (i.e. every participant completed both a pre and post-AGL perceptual task), and participants performed at ceiling in the post-AGL perceptual task. However, this could be attributed to practice effects: participants were already familiar with the perceptual task at the end of the AGL experiment. To prevent this, we divided participants into two groups, one group only completed the perceptual task before the artificial grammar learning task (perceptual initial group), and the other only completed it after (perceptual final group).

a. Materials and design

For the post-AGL perceptual task, we used the same stimuli from the pre-AGL perceptual task. In the following data, no participant completed both pre- and post-AGL perceptual tasks: using the same stimuli thus makes the results more comparable across the two groups. Participants heard two sounds (VC_1 – VC_2, or $V_1C_1V_2$ – $V_1C_2V_2$), and were asked to decide whether C_1 and C_2 were the same consonant. The two tokens in each pair were spoken by two different speakers, one female, one male, and were overlaid with multitalker babble at a signal to noise ratio of -10dB. The consonants tested were the following pairs: t-ʧ, s-ʃ, z-ʒ, b-f (control pair, non-alternating phonemes in the AGL task). Vowels were a, i, or u. Participants heard 12 random VC_1 – VC_2 and 36 $V_1C_1V_2$ – $V_1C_2V_2$ trials involving each pair of sounds combined with each vowel or pair of vowels, approximately half same, half different (total 192 trials, lasting about 15 minutes).

b. Procedure

Participants sat in front of a computer in a soundproof room, wearing noise-cancelling headphones. The experiment was conducted online using the Experigen experiment building platform (Becker & Levine 2010). Participants pressed the enter key on a keyboard to play each of the two tokens in an AX pair (e.g. [ata] by speaker 1, [ata] by speaker 2). They clicked buttons on screen with the mouse cursor to answer “Yes” or “No” to the question: “Did you hear the same consonant in both words?”. Each token could only be played once.

c. Participants

79 undergraduate students at UCLA participated in the post-AGL perceptual task (mean age = 22.00, SD = 5.99). 54 participants received course credit for their participation and 25 received 10USD in cash. Participants who received course credit were recruited through the UCLA Psychology Department Subject Pool, and paid participants through flyers on the UCLA campus.

Among all 156 participants who completed the AGL task, approximately half completed the AX task before the AGL task, and the other half after. The full repartition of participants was as listed in Table 10.

Table 10. *Total number of participants in AGL and AX condition (number of participants excluded from analysis in parentheses, exclusion criteria defined in results section)*

	Non-neutralizing /t/	Non-neutralizing /s/	Non-neutralizing /z/	Total
AX before AGL	27	24 (2)	26	77 (2)
AX after AGL	27 (2)	25	27	79 (2)
Total	54 (2)	49 (2)	53	156 (4)

d. Data analysis

Data was analyzed using linear mixed effect models. Just like in the baseline AX task, the dependent variable was participants' d' sensitivity index for each of the sound pairs. We first ran a model included Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f) and the Order in which participants completed perceptual and AGL tasks (perceptual initial, perceptual final) as fixed effects.

Since b-f was a control and words ending in b or f only appeared as non-alternating fillers in the AGL task (e.g. togab-togabi, savif-savifi), a second model excluded data from the b-f pair. This model included Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f), and Condition (baseline, neutralizing, non-neutralizing) as fixed effects.

In both models, random effect structure was kept maximal under condition of convergence (random intercepts and slopes for Participants).

e. Results

Four participants were excluded for reporting after the study that they had experienced speech or hearing impairments (2.6% of the data). Those were the same participants excluded from the analysis of the AGL task.

In our first model, we examined whether participants were more or less accurate in distinguishing the members in each of the tested sound pairs (t-ʧ, s-ʃ, z-ʒ, b-f) depending on whether they completed the perceptual task before or after the AGL task. Figure 25 shows participants' d' for each of the sound pairs in the two task orders. As before, the higher the d', the more participants were perceptually sensitive to the difference between two members of a sound pair, and those measures take into account response biases (in case participants preferred clicking one response button over the other).

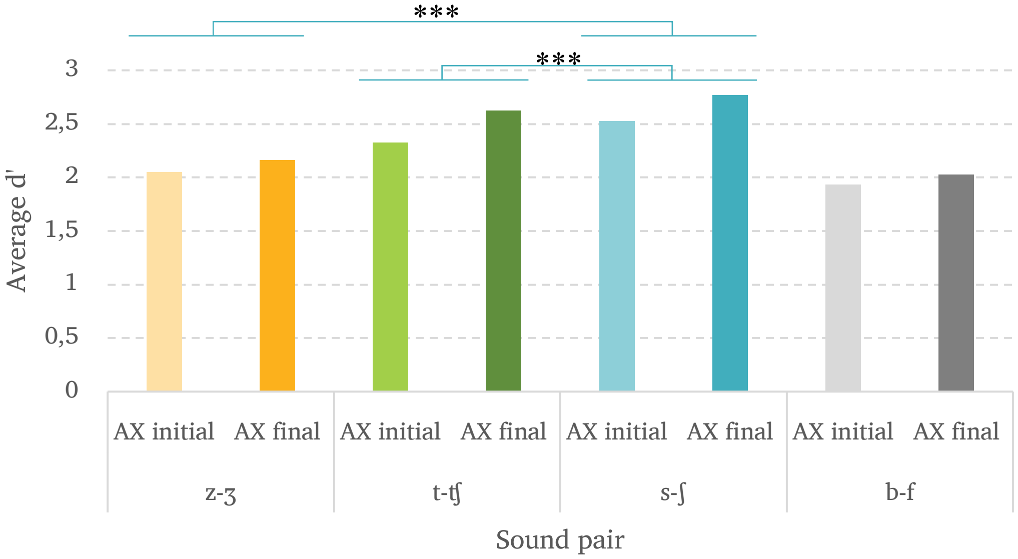


Figure 25*. d' scores depending on sound pair, English listeners perception before and after completing the AGL task*

We found no main effect of Order: overall, completing the perceptual task before or after the AGL did not significantly improve d'. Just as in the baseline perceptual results, we found a main effect of sound pair: t-ʧ and s-ʃ were more perceptually distinct than z-ʒ and b-f. Table 11 summarizes coefficients from this first model.

Table 11. *Coefficients in a model fitted to d', fixed effect of sound pair and experiment order (random intercepts and slopes for Participants)*

		Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
Intercept (b-f, AX initial)		1.93	0.10	444	19.20	<0.001	***
Order	AX final	0.09	0.14	148	0.65	0.52	
Sound pair	s-ʃ	0.59	0.09	444	6.26	<0.001	***
	t-ʧ	0.39	0.09	444	4.32	<0.001	***
	z-ʒ	0.12	0.10	444	1.22	0.23	
Interactions	s-ʃ : AX final	0.15	0.13	444	1.15	0.25	
	t-ʧ : AX final	0.21	0.13	444	1.63	0.10	
	z-ʒ : AX final	0.02	0.13	444	0.16	0.87	

Pairwise comparisons using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018) confirmed that b-f was not significantly different from z-ʒ in any task order. t-ʧ and s-ʃ both had higher d' values than b-f and z-ʒ, but t-ʧ and s-ʃ were not significantly different from one another. This is summarized in Table 12.

***Table 12**. Coefficients for pairwise comparisons of sound pairs from initial and final AX tasks (random intercepts and slopes for Participants)*

		Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
AX initial	b-f / z-ʒ	-0.12	0.10	444	-1.22	0.93	
	s-ʃ / z-ʒ	0.47	0.09	444	5.44	<0.0001	***
	t-ʧ / z-ʒ	0.28	0.08	444	3.42	0.02	*
	s-ʃ / t-ʧ	0.20	0.08	444	2.58	0.17	
AX final	b-f / z-ʒ	-0.14	0.09	444	-1.47	0.82	
	s-ʃ / z-ʒ	0.60	0.09	444	7.03	<0.0001	***
	t-ʧ / z-ʒ	0.46	0.08	444	5.79	<0.0001	***
	s-ʃ / t-ʧ	0.14	0.08	444	1.89	0.57	

The results from this first model can be schematized as:

s-ʃ, **t-ʧ** > **z-ʒ**, **b-f**

where > means "more perceptually distinct than". Setting aside again the case of b-f, this order is again consistent with the functional load ranking according to both entropy and minimal pair count. What we were most interested in in this comparison, though, was (1) whether post-AGL perceptual distinctiveness would be different for contrasts with different functional loads and (2) whether the type of alternation learned during the AGL (neutralizing or non-neutralizing) mattered, depending on the functional load of the contrasts. To answer these questions, we fitted the data to a second model. The first model showed that our control pair, b-f, showed no significant improvement in perceptual distinctiveness after the AGL task. Since b and f were present as non-alternating filler endings (e.g. togab-togabi, savif-savifi) in the AGL task, participants who completed the final AX task had received more exposure to these sounds than those who completed the initial AX task. Therefore, additional exposure alone was not sufficient to improve the perceptual distinctiveness of these sounds. However, the other sound pairs (z-ʒ, t-ʧ and s-ʃ) were involved in either neutralizing or non-neutralizing alternations in the AGL task. In our second model, we excluded data from the b-f pair (which only had an AX initial/AX final distinction), and compare the baseline d' for the three other pairs (AX initial) to their d' after an AGL with either a neutralizing or non-neutralizing alternation. Figure 26 shows these results.

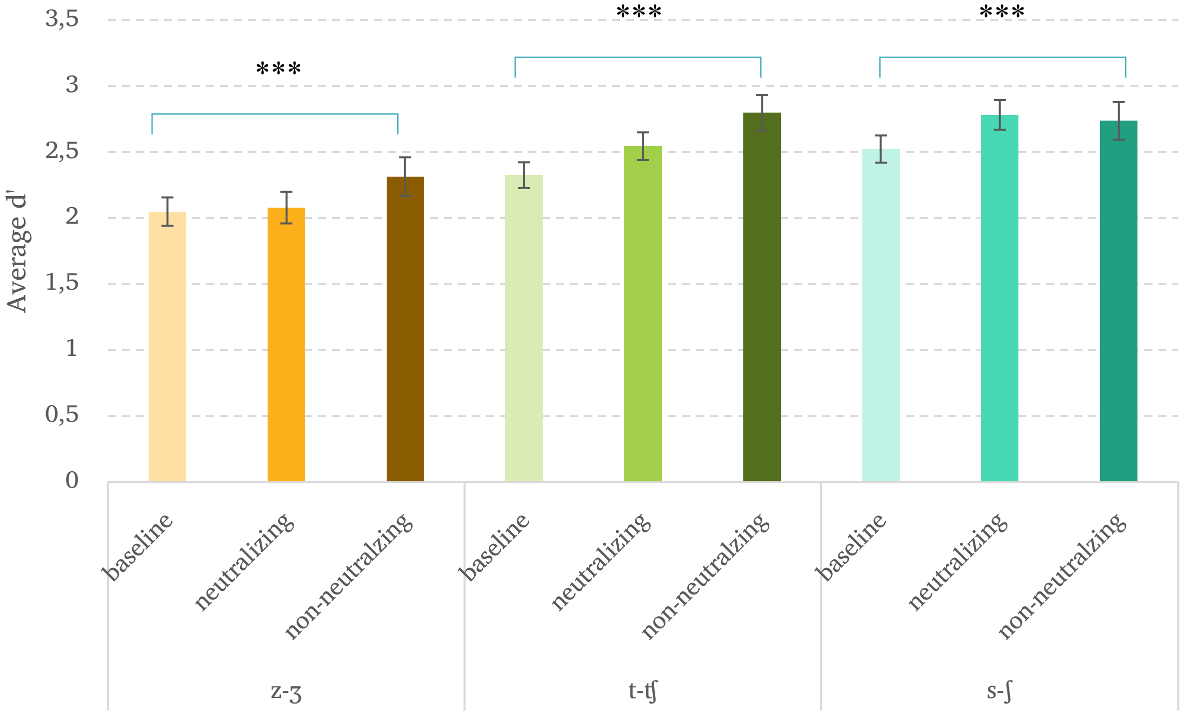


Figure 26. *d' index depending on sound pair, English listeners perception before and after completing the AGL task with neutralizing or non-neutralizing alternations*

Table 13 summarizes all coefficients from our second model.

***Table 13**. Coefficients in a model fitted to d', fixed effects of sound pair and alternation (random intercepts and slopes for Participants)*

		Estimate	Std. Error	DF	t-value	p-value	
Intercept (s-ʃ, baseline)		2.52	0.10	292	24.47	<0.001	***
Alternation	neutralizing	0.17	0.15	292	1.11	0.27	
	non-neutralizing	0.39	0.18	292	2.24	0.03	*
Sound pair	t-ʧ	-0.20	0.08	292	-2.59	0.01	**
	z-ʒ	-0.47	0.09	292	-5.51	<0.001	***
Interactions	t-ʧ : neutralizing	0.11	0.13	292	0.82	0.41	
	t-ʧ : non-neutralizing	-0.04	0.18	292	-0.24	0.81	
	z-ʒ : neutralizing	-0.11	0.14	292	-0.77	0.44	
	z-ʒ : non-neutralizing	-0.18	0.19	292	-0.94	0.35	

We found a main effect of alternation: overall, the perceptual distinctiveness of a sound pair improved after learning a non-neutralizing alternation, but not after learning a neutralizing alternation. This is confirmed by a post-hoc comparison of means using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018)): as there were no significant interactions, we compared means for all sound pairs (Table 14).

***Table 14**. Coefficients for pairwise comparisons of baseline perceptual distinctiveness and final distinctiveness after learning a neutralizing or a non-neutralizing alternation*

	Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
baseline vs. neutralizing alternation	-0.17	0.13	292	-1.29	0.40	
baseline vs. non-neutralizing alternation	-0.32	0.14	292	-2.33	0.05	*
neutralizing vs. non-neutralizing alternation	-0.15	0.07	292	-2.15	0.08	

Just as in the previous results, we also found a main effect of sound pair: s-ʃ was more perceptually distinct than t-ʧ and z-ʒ. A post-hoc comparison of means using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018)) confirms this, and furthermore that t-ʧ was more distinct than z-ʒ (Table 15).

***Table 15**. Coefficients for pairwise comparisons of perceptual distinctiveness for the three test sound pairs in all AX tasks*

	Estimate	**Std. Error**	**DF**	**t-value**	**p-value**	
s-ʃ / z-ʒ	0.57	0.07	292	8.42	<0.0001	***
z-ʒ / t-ʧ	-0.39	0.06	292	-6.10	<0.0001	***
s-ʃ / t-ʧ	0.18	0.06	292	2.86	0.01	*

These results can be schematized as:

s-ʃ > **t-ʧ** > **z-ʒ**

where > means "alternations involving [sound pair on the left] were learned better than". This order is fully consistent with the functional load ranking according entropy, but not minimal pair count (in Figure 16 repeated below).

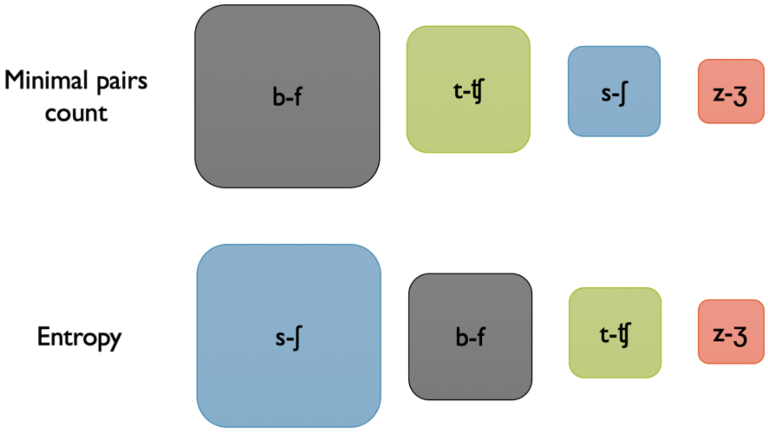


Figure 16 (repeated). *Functional load rank ordering of test contrasts in the two measures*

Earlier, we left aside the case of b-f, a rather high-functional-load contrast in both measures. The final comparison of AX results can only tell us that the perceptual distinctiveness of this pair did not improve after exposure to these sounds in the AGL task. However, the fact that the d' did not change for this pair serves as a control to show that being involved in a non-neutralizing alternation improves perceptual distinctiveness. The effect of functional load, for the three other pairs, manifests in two ways: (1) higher functional load contrasts are easier to tell apart in noise and (2) learning a phonological alternation that does not reduce functional load (non-neutralizing) improves their perceptual distinctiveness, but learning one that reduces functional load (neutralizing) does not.

10. Influence of perceptual task on AGL

Distinctiveness improved after the AGL task for the test sound pairs if they were involved in a non-neutralizing alternation, but not if they were involved in a neutralizing alternation, and not for the control pair b-f. We attribute this to a difference in whether the functional load of these contrasts was manipulated in the AGL task: whether the sound pairs examined in the perceptual task were involved in an alternation, and whether the alternation was neutralizing or not.

One could argue that learning the test sound pairs in alternations in the AGL simply brought participants' attention to these sounds, and thus increased perception. The distinctiveness of b-f, then, did not improve because despite being present in the AGL stimuli, they did not alternate with each other in the artificial languages. If better performance in the perceptual task depended only on the AGL increasing awareness of the test sound pairs, could it be the case that performance in the AGL was also affected by whether it was preceded by an AX task that brings attention to the alternating contrasts? If that were the case, our argument for the role of functional load in causing the shift in perceptual distinctiveness would be much weakened, as any change in performance could then be attributed to how much attention was paid to the sound pairs during the two tasks. The fact that there was a difference between cases where participants learned a neutralizing alternation vs. a non-neutralizing runs against this explanation, but we can provide more evidence by analyzing the AGL task order.

To check whether this was the case, we compared the AGL results by sound pair from participants who completed it before or after the perceptual task (Figure 27). To this end, we used a mixed effects logistic regression with Order (AGL first vs. AGL after) and Sound Pair (t-ʧ, z-ʒ, s-ʃ, b-f) and their interaction as fixed effects. Random effect structure was kept maximal under condition of convergence (random intercepts and slopes for Participants and random intercepts for Items, Table 16).

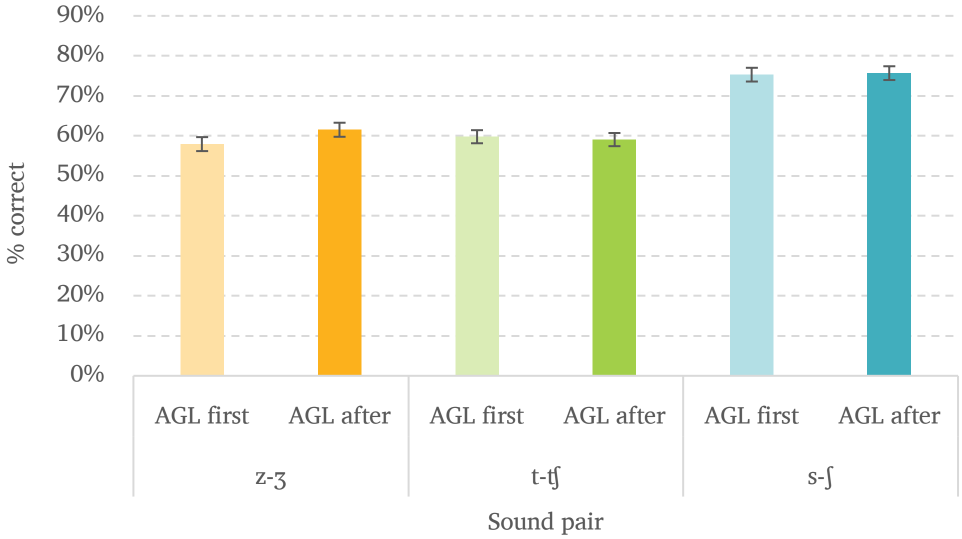


Figure 27*. Percentage of correct answers per sound pair, depending on whether the AGL task was completed before or after the AX task*

Table 16*. Coefficients in a model fitted to correct answers in the AGL task, fixed effects of sound pair and experiment order (random intercepts and slopes for Participants, random intercepts for Items)*

		Estimate	**Std. Error**	**z-value**	**p-value**	
Intercept (s-ʃ, AGL first)		1.35	0.21	6.38	<0.001	***
Order	AGL after	-0.03	0.17	-0.15	0.88	
Sound pair	t-ʧ	-1.10	0.25	-4.35	<0.001	***
	z-ʒ	-0.82	0.26	-3.10	<0.01	**
Interactions	t-ʧ : AGL after	-0.02	0.11	-0.21	0.83	
	z-ʒ : AGL after	0.22	0.11	2.04	0.04	*

We found a significant interaction: the difference between the AGL results before or after the AX task was greater for the z-ʒ pair than for the s-ʃ pair (this was not the case for the t-ʧ pair). However, a post-hoc comparison of means using least square means for multiple comparisons (lsmeans package in R(R Core Team 2018)) showed that the learning for each sound pair did not differ significantly regardless of whether the alternation involving them was learned before or after completing the perceptual task (Table 17). The significant interaction in the model was then driven by the fact that means mostly differed by sound pair, a significant main effect in the model, and that performance for the s-ʃ pair was much better than for the z-ʒ pair.

Table 17*. Coefficients for pairwise comparisons of AGL results before and after the AX task for each sound pair*

	Estimate	**Std. Error**	**DF**	**t-value**	**p-value**
s-ʃ, AGL first vs. after	0.03	0.17	Inf	0.15	1.00
z-ʒ, AGL first vs. after	-0.20	0.16	Inf	-1.20	0.84
t-ʧ, AGL first vs. after	0.05	0.16	Inf	0.30	1.00

This allows us to conclude that it was the learning of alternations in the AGL task which induced changes in perceptual distinctiveness, and that it was actually the difference in phonological status of the sound pairs, rather than a general effect of attention, that cause the change. For participants who completed the AX task before the AGL, having their attention brought to the test sound pairs in the AX task was not sufficient to improve their phonological learning in the AGL. For those who completed the AX task after the AGL, the phonological learning of non-neutralizing alternations in the AGL improved in the AX task the perceptual distinctiveness of the sound pairs involved in the alternation.

11. Learning in all conditions

Lastly, we checked that the differences in learning were not due to an overall difference in performance across the 3 artificial language conditions. Again, we used a mixed effects logistic regression with Order (AGL first vs. AGL after) and Language (non-neutralizing /s/, non-neutralizing /t/ non-neutralizing /z/) and their interaction as fixed effects. Random effect structure was kept maximal under condition of convergence (random intercepts and slopes for Participants and random intercepts for Items). Figure 28 shows these results.

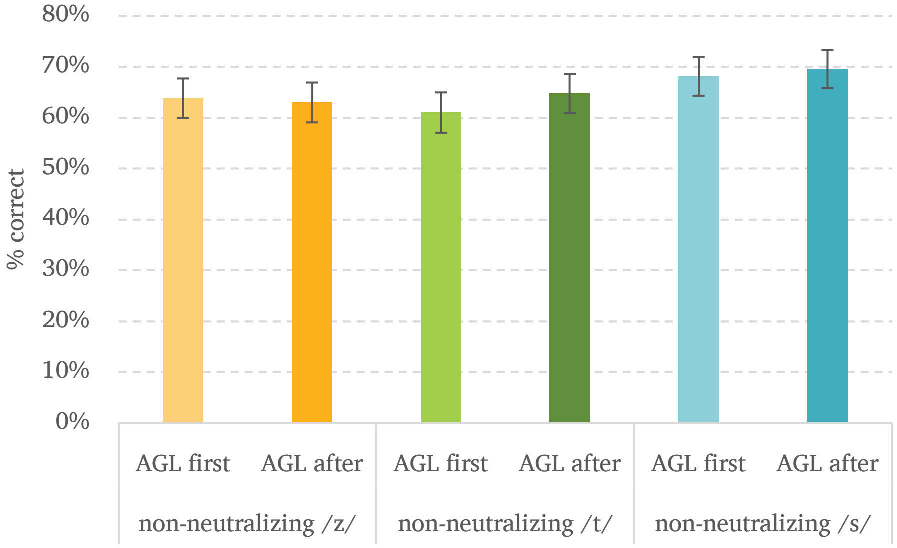


Figure 28. *Percentage of correct answers per language condition, depending on whether the AGL task was completed before or after the AX task*

Learning was not significantly different across the three artificial languages, and experiment orders (see Table 18 for detailed coefficients). No one language was easier to learn than the others: as reported in a previous section, only sound pair mattered for how well participants learned the alternation, with alternations involving the high functional load pair s-ʃ being learned better than those involving t-ʧ or z-ʒ.

***Table 18**. Coefficients in a model fitted to correct answers in the AGL task, fixed effects language condition and experiment order (random intercepts and slopes for Participants, random intercepts for Items)*

		Estimate	**Std. Error**	**z-value**	**p-value**	
Intercept (non-neutralizing /s/, AGL first)		0.76	0.14	5.23	<0.001	***
Order	AGL after	0.07	0.21	0.33	0.74	
Language	non-neutralizing /t/	-0.24	0.21	-1.15	0.25	
	non-neutralizing /z/	-0.13	0.20	-0.66	0.51	
Interactions	non-neutralizing /t/:AGL after	0.05	0.29	0.17	0.87	
	non-neutralizing /z/:AGL after	-0.10	0.29	-0.35	0.72	

12. Chapter conclusion

In this series of experiments, we tested the relationship between functional load, perceptual distinctiveness, and phonological learning. In a first perceptual AX task, we established a baseline of perceptual distinctiveness for four pairs of segmental contrasts in English. The lowest functional load pair according to both measures, z-ʒ, was also the most difficult to tell apart. Higher functional load s-ʃ and t-ʧ were both more perceptually distinct, but at a comparable level. The b-f pair, which carries rather high functional load in both measures, had surprisingly low perceptual distinctiveness. This is however consistent with previous confusion matrix data, and we can only hypothesize that this is due to greater acoustic similarity. The ranking in terms of perceptual similarity in the AX task was consistent with both measures if we set aside the case of b-f: higher functional load contrasts were more perceptually distinct.

Results from this first AX task could not exclude the possibility that all the differences we observed were only due to acoustic affects. We then turned to an artificial grammar learning task to assess the phonological learning of alternations involving sound pairs from contrasts with different functional load. We examined two aspects of this: (1) whether alternations involving a sound pair from a high-functional-load contrast would be learned better, and (2) whether neutralizing alternations, which reduce the functional load of a contrast in the artificial language, are learned differently from non-neutralizing ones. We partially replicate Yin and White's (2018) result that if the alternations in the artificial language do not induce homophones, neutralizing and non-neutralizing alternations are learned to the same extent. However, we found that this was not true for all contrasts: for the s-ʃ pair, the neutralizing alternation was actually learned better.

Overall, the results from the AGL task rank the s-ʃ pair above t-ʧ and z-ʒ in terms of learning alternations. Taken with the first AX task results, which set s-ʃ and t-ʧ apart from z-ʒ, these results are more consistent with the entropy measure of functional load (s-ʃ > t-ʧ > z-ʒ) than the minimal pair count measure (t-ʧ > s-ʃ > z-ʒ).

This is further supported by comparing the initial AX results to the same task completed after the AGL task. Beside the difference by sound pair (s-ʃ > t-ʧ > z-ʒ) where higher functional load led to better perceptual distinctiveness, we found that the perceptual distinctiveness of a sound pair improved after learning a non-neutralizing alternation, but not after learning a neutralizing alternation. The control pair (b-f) that appeared in the AGL but was not involved in any alternations did not show significant improvement in perceptual distinctiveness after the AGL task. This means that phonological learning was key in inducing the change in perception. The fact that sound pairs involved in neutralizing alternations also failed to improve in perceptual distinctiveness despite being learned to a similar extent in the AGL task suggests that it was specifically neutralization that impeded the improvement. Following Boomershine et al.'s (2008) findings that sounds that do not contrast phonemically in L1 are more perceptually similar, one might have expected that non-neutralizing alternations would be the ones impeding perceptual improvement, since members of the sound pair are not separate phonemes in non-neutralizing alternations, whereas they are in neutralizing alternations. However, we can actually relate our results to Yin and White's (2018) finding that speakers track homophony in an AGL. Neutralizing alternations cause a loss of information in a way that non-neutralizing alternations do not. Our participants were sensitive to two types of functional load. First, the functional load of sound pairs in their native language: higher functional load contrasts were more perceptually distinct, and alternations involving them were learned better. Second, the functional load of these same sound pairs in the artificial language they were taught: in the case of neutralizing alternations, lexical contrasts could be lost through the alternation of these sounds. As to non-neutralizing alternations, despite a change in phonemic status compared to L1, there is actually no information lost in the artificial language itself, since there is still a one-to-one mapping between allophone and phoneme in the alternating context. This difference in whether the phonological alternation causes a change in the information structure of the language was then reflected in the change (or lack thereof) in the perceptual distinctiveness of the sound pairs tested. For neutralizing alternations, participants were taught that the difference between the two sounds of a pair was less informative than in their native language, since it neutralizes in some positions. For non-neutralizing alternations, participants were taught that the two sounds mapped unambiguously to a single phoneme, and that they provided clues to the following phonological context (_[i] or not). This difference in informativity might explain why perception only improved after learning non-neutralizing alternations, and not neutralizing ones.

The order of the experiments (AX task before AGL or AX task after AGL) only affected the results of the AX task, not those of the AGL. This confirms that it was phonological learning that affected perception, while the perceptual task did not affect learning. Furthermore, learning did not differ significantly in the three artificial languages, so we can really attribute the differences observed to the sound pairs tested and the alternation type that were taught.

Chapter V. Conclusion

13. Summary of findings

The aim of this series of simulations, corpus studies and experiments was to look for synchronic evidence of the functional load hypothesis. We conducted this investigation following two main axes: (1) looking for effects of functional load in the current state of a language, and (2) looking for effects of functional load in learning an artificial language. But first, as there are two main methods of calculating functional load that yield different results, we needed to compare them. We used a series of simulations and corpus studies to examine the conceptual differences between these measures, the variation in entropy and the count of minimal pairs.

a. Simulations and corpus studies: comparison of functional load measures

While our simulations with toy corpora showed that the two measures of functional load have the mathematical potential of being very different, we also found that they are in fact well correlated in real language corpora. However, despite this correlation, the two measures did not rank sound contrasts in the same order, and the same contrast can have a high functional load according to one measure, and a low one according to the other. Later, we relied on this difference in rank ordering to compare the two measures using experimental results.

Our first line of investigation consisted in looking for a relationship between functional load, taken as a metric of lexical contrastiveness, and perception. Using data from pre-existing confusion matrices for English (Wang & Bilger, 1973), we correlated this perceptual confusability to functional load in both English and French. If functional load impacts perception within a language, we expected the confusability data to correlate with functional load in English, but not in French.

While we found a significant correlation between perceptual confusability and functional load according to both measures within a language (English), the correlation ran in opposite directions for the two measures. This did not straightforwardly support our hypothesis that high functional load contrasts are easier to tell apart, but our control case (French) shows that these correlations could not be attributed merely to a universal preference for acoustic distinctiveness in lexical contrasts.

The opposite directions of the correlation with the two measures of functional load could have been due to the differences in rank ordering of the contrasts according to the two measures of functional load (which we found during our corpus study). That is, it is possible that there is a consistent relationship between functional load and perception, but one of the measures is more suited to capturing this effect. However, there was a possibility that this apparent contradiction stemmed from the nature of the task used to establish the confusion matrices we relied on. To address this confound, we designed a perceptual discrimination experiment.

b. Experiment 1 (part 1): Functional load and baseline perceptual distinctiveness in L1

To establish a baseline of the discriminability of some segmental contrasts for English listeners, we conducted a perceptual discrimination task. We tested the discriminability in noise of four sound pairs, z-ʒ, s-ʃ, t-ʧ, and b-f, which all had different functional load according to the two measures in English. Additionally, the two measures ranked them differently (see Figure 3, repeated)

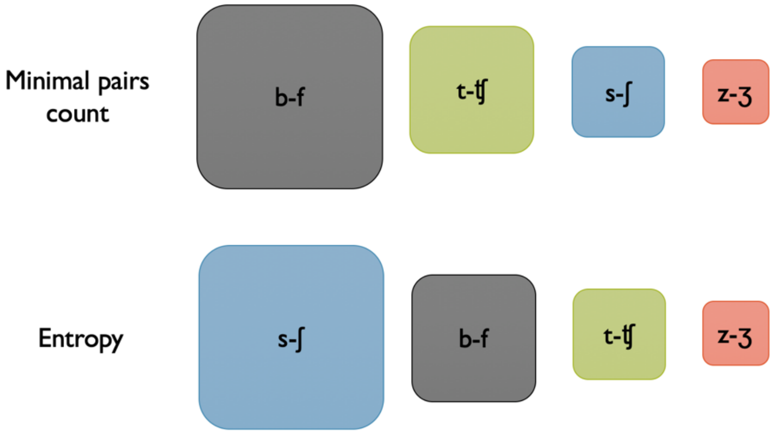


Figure 16 (repeated) . *Functional load rank ordering of test contrasts in the two measures*

The lowest functional load contrast according to both measures, z-ʒ, also had the lowest perceptual discriminability. The higher functional load contrasts, s-ʃ and t-ʧ, were only marginally different, and there was thus no clear-cut evidence in support of either of the two measures against the other. Additionally, b-f should have a relatively high functional load, but was as difficult to tell apart as z and ʒ.

The relatively low perceptual distinctiveness of b-f is consistent with confusion matrix data from previous studies. We could conjecture that despite their relatively high functional load in English, b and f are just acoustically too similar, and this explains why they were not significantly different from the lowest functional load z-ʒ in the baseline perceptual task. This forces us however to consider that the differences we found among the other sound pairs might also be due to pure acoustic (dis/)similarity.

Our corpus studies and this perceptual discrimination experiment allowed us to test hypotheses about functional load and its role in the current, state of the English language. The central claim of the functional load hypothesis, however, is that functional load plays a role in the evolution of a language.

Diachronic corpus studies (Todd 2012, Wedel et al. 2013) have found evidence to support this. In our study, we wanted to see if functional load from one's native language also impacted the learning of phonological patterns in an L2. To this end, we also conducted an artificial grammar learning experiment. Would alternations involving high functional load sound pairs be easier or more difficult to learn? Does it matter whether the alternations are neutralizing or not?

The artificial grammar learning experiment (AGL) would also allow us to test the effect of learning an alternation on perceptual distinctiveness. If perceptual distinctiveness depends on the contrastiveness (and thus functional load) of the sounds compared (Boomershine et al. 2008), phonological learning might well induce a change in perception. We did indeed find such a change after the AGL task.

c. *Experiment 2: Artificial grammar learning task*

In the artificial grammar learning task, we tested whether alternations involving sound pairs with different functional loads were learned differently, and whether it mattered that the alternation was neutralizing or not. Three conditions across participants allowed us to teach alternations involving each of three sound pairs (t-ʧ, z-ʒ, s-ʃ) in a neutralizing and a non-neutralizing alternation.

We partially replicated Yin and White's (2018) result that neutralizing alternations (that create no homophones) were not learned differently from non-neutralizing ones involving the same sound pairs: this is true for t-ʧ and z-ʒ, but not s-ʃ, where the neutralizing alternation was actually learned better.

Overall, the results from the AGL task show that alternations involving the s-ʃ pair are easier to learn than those involving t-ʧ and z-ʒ. Taken together with the results of the baseline perceptual task that ranked s-ʃ and t-ʧ above z-ʒ, this gave us a tentative ranking of:

s-ʃ > **t-ʧ** > **z-ʒ**

which is consistent with the entropy measure but not the minimal pair count.

If contrastiveness affects perceptual distinctiveness, and learning phonological alternations involves learning new functional loads for native contrasts, we should expect a change in the perceptual distinctiveness of the sound pairs involved after the AGL task. We tested this using another perceptual discrimination task after the AGL.

d. *Experiment 1 (part 2): Functional load and post-AGL perceptual distinctiveness*

Half of the participants in the AGL experiment completed the perceptual task before the AGL, and half after. We compared the perceptual results of these two groups, whose performance in the AGL did not significantly differ.

Beside the significant difference by sound pair (s-ʃ > t-ʧ > z-ʒ) where higher functional load led to better perceptual distinctiveness, we found that the perceptual distinctiveness of a sound pair improved after learning a non-neutralizing alternation, but not after learning a neutralizing alternation.

The b-f contrast served as a control: these sounds appeared in the AGL fillers, but were not involved in any alternations. The perceptual distinctiveness of this pair did not show significant improvement after the AGL task. This means that phonological learning was key in inducing the change in perception. The fact that sound pairs involved in neutralizing alternations also failed to improve in perceptual distinctiveness despite being learned to a similar extent in the AGL task suggests that it was specifically neutralization that impeded the improvement. A possible future analysis would involve correlating individual participants' learning in the AGL to their performance in terms of perceptual distinctiveness. As completing the perceptual task first did not impact group learning in AGL, we should predict a difference mainly for participants who completed the perceptual task after the AGL, and a negative correlation between success in learning neutralizing alternations and perceptual distinction of the neutralized sound pairs.

Boomershine et al.'s (2008) findings that non-phonemic contrasts in L1 are more perceptually similar could lead us to expect non-neutralizing alternations to impede perceptual improvement, rather than neutralizing ones. Indeed, members of a sound pair in non-neutralizing alternations were not separate phonemes, whereas they were in neutralizing alternations. We conjecture that our result related more to Yin and White's (2018) finding that speakers track homophony, and thus the information structure, in the artificial language. Although the neutralizing alternations in our experiment did not create homophones, they did cause a loss of information in a way that non-neutralizing alternations do not, and listeners were sensitive to this difference. In the case of neutralizing alternations, lexical contrasts could be lost through the alternation of the sounds involved (and even if there were no actual homophones created in our stimuli, many word pairs became less distinct, e.g. /kotaʃ/ and /fotas/ become [kotaʃi] and [fotaʃi], now differing by one segment instead of two). As to non-neutralizing alternations, despite a change in phonemic status compared to participants' L1, there is actually no information lost in the artificial language itself, since there is still a one-to-one mapping between allophone and phoneme in the alternating context.

This difference in how alternations impacted the information structure of the language was then reflected in the change (or lack thereof) in the perceptual distinctiveness of the sound pairs tested. In neutralizing alternations, participants learned that the difference between the two sounds of a pair was less informative than in their native language, since it neutralizes in some positions. In non-neutralizing alternations, participants learned that each of the two sounds mapped unambiguously to a distinct phoneme, and that they provided clues to the following phonological context (_[i] or not). This difference in informativity could explain why perception only improved after learning non-neutralizing alternations, and not neutralizing ones.

Our experimental results showed that participants were sensitive to two types of functional load. First, the functional load of sound pairs in their native language: higher functional load contrasts were more perceptually distinct, and alternations involving them were learned better. Second, the functional load of phonemes in the artificial language: learning alternations means learning that the test sound pairs carried more or less information in the new language, and this impacted the perceptual distinctiveness of the sound pairs.

14. Directions for future research

Our findings provided some evidence towards a synchronic version of the functional load hypothesis. High-functional-load contrasts are better perceived, and learned better in an artificial language. Furthermore, learning an artificial language with a different information structure for native contrasts can affect perception. The perceptual advantage could explain how high functional load contrasts are preserved across generations of speakers. The learning advantage suggests that speakers track the amount of information conveyed by each contrast in their native lexicon, and they also track this in a new language. This information can then change listeners' perception of these contrasts.

Of course, the scope of the dissertation is limited, and there remain issues we did not have clear explanations for. We discuss some of them in the following sections.

a. *Perceptual distinctiveness: disentangling purely acoustic effects from functional load effects*

In our perceptual discrimination experiment, we found the b-f contrast, despite having rather a high functional load in English, was difficult to tell apart perceptually. This was consistent with previous confusion matrix data, and probably results from pure acoustic similarity. This is probably not the only sound pair for which this is the case. A problem arises then: how to disentangle pure acoustic effects from functional-load-induced perceptual effects? In our study, we side-stepped this issue since we mainly needed to compare the changes in perceptual distinctiveness following the artificial grammar experiment. However, it might be possible to single out a perceptual effect of functional load using cross-linguistic comparison.

This could involve running the same perceptual experiment (possibly using synthesized, or partly synthesized stimuli) with participants with different language backgrounds, preferably from widely different language families. An approximation of pure acoustic distinctiveness could be obtained by averaging participants' performance across language groups: since this should only depend on the human (or mammalian) auditory system, it should be universal across languages. The difference between each language group's performance, and their difference to that mean, should help estimating a language specific functional-load-induced perception effect.

b. Learning neutralizing alternations with different functional loads

Another question that would benefit from more evidence is whether functional load can impact the learning of neutralizing vs. non-neutralizing alternations. We had found that for the high-functional-load contrast s-ʃ, the neutralizing alternation was learned better than the non-neutralizing one. We do not know whether this effect is robust, and whether it might be contrast-specific or language-specific. Ideally, conducting the AGL task using different sets of contrasts, different types of phonological processes and with participants from various language backgrounds for which there is a large corpus from which to compute functional load would provide the evidence to assess this.

If this effect were robust cross-linguistically, it would lead us to conjecture that in our experiment the effect of functional load might be different in the AGL and in shaping perception.

c. Language learning and evolution of perception

There is a plausible explanation as to why the s-ʃ neutralizing alternation was learned better. Despite the loss of contrast in the neutralizing environment, /s/ and /ʃ/ remained separate phonemes (just as in our participants' native language, English) when they were involved in a neutralizing alternation. However, in non-neutralizing alternations, they became allophones of the same phoneme /s/. As they form a high-functional-load contrast in their native language, listeners might have been unwilling to assign them to a single phoneme, even in an artificial L2.

However, in our interpretation for the change in perception (or lack thereof) after the AGL task, we argued that perceptual distinctiveness did not improve after learning neutralizing alternations because they caused a loss of information in the artificial L2. At first glance, these arguments can seem contradictory. However, if the effect of functional load is different on learning and on perception, these could both be true.

In learning, adult listeners initially reference the functional load of contrasts in their L1, and thus might prefer to preserve the phonemic status of L1 sounds, even in the L2. However, functional effects on perception stem from the phonological and lexical information of all the languages a listener speaks, and thus the learning of an L2, even artificial, could modify their perception. Adding to this, for participants who completed the perceptual task after the AGL, the perceptual task immediately followed the AGL, which could have encouraged furthermore the reference to the artificial-language-internal functional load, for which neutralizing alternations caused a loss of information, but non-neutralizing ones did not.

Another way to investigate the questions we attempted to answer with our series of experiments would be to assess phonological learning and look for perceptual changes in real L2 learners. For speakers learning an L2 where some native contrasts neutralize, do they learn the neutralization as well as other phonological processes in the language? This could be measured in terms of accuracy, but also in terms of how early neutralization is acquired in the learning process. And for more or less advanced learners, does this also shape their perception of the sounds involved in neutralization? This could be examined using perceptual discrimination tasks. This entire situation would then be compared to another group of learners learning a language where the same sounds are allophones of a same phoneme.

Appendix I. Functional load according to entropy and minimal pairs for consonant contrasts in English

Functional load according to entropy and minimal pair count for consonant contrasts in English, using the IPHOD corpus (Irvine Phonotactic online Dictionary, Vaden, Halpin & Hickok, 2009). For entropy, we use word type frequency, and for minimal pairs, the count of minimal pairs relativized to the number of possible minimal pairs. Both measures were computed using Phonological Corpus Tools (Hall et al., 2018), and were relativized to the size of the corpus. Sound pairs tested in our experiments are highlighted.

sound pair	FL entropy	FL minpairs	sound pair	FL entropy	FL minpairs	sound pair	FL entropy	FL minpairs	sound pair	FL entropy	FL minpairs
b-d	0.00125	0.004999	ð-ʤ	5.40E-05	0.009198	f-ʤ	0.000383	0.019558	g-v	0.000259	0.007708
b-ð	7.67E-05	0.013692	d-f	0.000807	0.024308	f-s	0.001069	0.003226	g-z	0.000181	0.008983
b-ʤ	0.000526	0.000599	d-g	0.000756	0.019872	f-ʃ	0.000452	0.010962	g-ʒ	0	0.010562
b-f	0.001181	0.012854	d-s	0.00148	0.073227	f-ʧ	0.000427	0.033744	g-θ	0.000153	0.003782
b-g	0.001011	0.015127	d-ʃ	0.000606	0.026634	f-v	0.000365	0.016938	k-b	0.001548	0.004879
b-s	0.001198	0.013823	d-ʧ	0.000617	0.006312	f-z	0.000243	0.013207	k-d	0.00155	0.016926
b-ʃ	0.000598	0.016276	ð-ʧ	6.25E-05	0.036087	f-ʒ	0	0.008002	k-ð	0.000138	0.018922
b-ʧ	0.000523	0.020725	d-v	0.000657	0.004986	f-θ	0.000132	0.017705	k-ʤ	0.000525	0.01655
b-v	0.000411	0.006878	d-z	0.004142	0.004073	g-ð	5.94E-05	0.01057	k-f	0.001265	0.015748
b-z	0.000262	0.018608	ð-z	6.36E-05	0.003868	g-ʤ	0.000275	0.000975	k-g	0.001154	0.016208
b-ʒ	8.10E-06	0.011943	d-ʒ	2.80E-05	0.01444	g-f	0.000637	0.011776	k-s	0.001607	0.008962
b-θ	0.000247	0.007226	ð-ʒ	5.40E-06	0.033275	g-s	0.000673	0.014495	k-ʃ	0.000709	0.014768
d-ð	0.000145	0.010055	d-θ	0.000346	0.018786	g-ʃ	0.000406	0.023253	k-ʧ	0.000673	0.000513

sound pair	FL entropy	FL minpairs
d-ʤ	0.000552	0.017623
k-z	0.000609	0.002993
k-ʒ	2.26E-05	0.020687
k-θ	0.00028	0.019081
p-b	0.00133	0.003888
p-d	0.001152	0.010949
p-ð	0.000111	0.000585
p-ʤ	0.00044	0.023668
p-f	0.001081	0.007733
p-g	0.000785	0.018913
p-k	0.002175	0.007421
p-s	0.001321	0.018595
p-ʃ	0.000598	0.022225
p-t	0.001759	0.01014
p-ʧ	0.000586	0.013339
p-v	0.00044	0.006035
p-z	0.000421	0.003203
f-ð	7.09E-05	0.012571
p-ʒ	8.10E-06	0.038352
p-θ	0.000239	0.004943
s-ð	9.20E-05	0.000804
s-ʤ	0.000531	0.014446
s-ʃ	0.000666	0.007433
s-ʧ	0.000562	0.000779
s-v	0.000546	0.007229
s-z	0.000612	0.009828
s-ʒ	2.53E-05	0.02968
ʃ-ð	6.19E-05	0.027548
ʃ-ʤ	0.00024	0.010774
ʃ-ʧ	0.000322	0.000693
ʃ-v	0.000162	0.000325
ʃ-z	0.000125	0.008939
ʃ-ʒ	1.62E-05	0.011077
t-b	0.001413	0.010952
g-ʧ	0.00031	0.008027
t-d	0.00188	0
t-ð	0.000166	0.011737
t-ʤ	0.000629	0.020603
t-f	0.00104	0.004835
t-g	0.000889	0.016945
t-k	0.002332	0.006093
t-s	0.003087	0.003702
t-ʃ	0.000733	0.020495
t-ʧ	0.000687	0.001595
t-v	0.000572	0.01052
t-z	0.000961	0.015964
t-ʒ	3.28E-05	0.023104
t-θ	0.000359	0.011591
ʧ-ʤ	0.000226	0.014782
v-ð	0.000307	0.001435
v-ʤ	0.00026	0.000439
k-v	0.000565	0.002929
v-ʧ	0.000179	0.025409
v-z	5.33E-05	0.001978
v-ʒ	1.08E-05	0
z-ʤ	0.000245	0.005713
z-ʧ	0.000259	0.003584
z-ʒ	2.36E-05	0.008901
ʒ-ʤ	1.18E-05	0.000966
ʒ-ʧ	5.40E-06	0.019153
θ-ð	1.35E-05	0.004749
θ-ʤ	9.82E-05	0.025275
θ-s	0.000274	0.001499
θ-ʃ	9.76E-05	0.021848
θ-ʧ	0.000127	0.011792
θ-v	7.06E-05	0.020934
θ-z	0.000155	0.005681
θ-ʒ	2.70E-06	0.012637

Appendix II. Experiment stimuli

1. AX perceptual task

perceptual task: VC		
at	aʒ	ab
it	iʒ	ib
ut	uʒ	ub
atʃ	as	af
itʃ	is	if
utʃ	us	uf
az	aʃ	
iz	iʃ	
uz	uʃ	

perceptual task: VCV			
ata	aza	asa	aba
ati	azi	asi	abi
atu	azu	asu	abu
ita	iza	isa	iba
iti	izi	isi	ibi
itu	izu	isu	ibu
uta	uza	usa	uba
uti	uzi	usi	ubi
utu	uzu	usu	ubu
atʃa	aʒa	aʃa	afa
atʃi	aʒi	aʃi	afi
atʃu	aʒu	aʃu	afu
itʃa	iʒa	iʃa	ifa
itʃi	iʒi	iʃi	ifi
itʃu	iʒu	iʃu	ifu
utʃa	uʒa	uʃa	ufa
utʃi	uʒi	uʃi	ufi
utʃu	uʒu	uʃu	ufu

Participants heard two sounds (VC_1 – VC_2, or $V_1C_1V_2$ – $V_1C_2V_2$), and were asked to decide whether C_1 and C_2 were the same consonant. The two tokens in each pair were spoken by two different speakers, one female, one male, and were overlaid with multitalker babble (Wilson et al. 2007) at a signal to noise ratio of -10dB.

Exposure: test items								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
kɛfos	kɛfoʃi	s	dɛbiʃ	dɛbiʃi	ʃ	ʒifot	ʒifoʧi	t
zɛkus	zɛkusi	s	vuzɛʃ	vuzɛʃi	ʃ	zuvot	zuvoʧi	t
zɛgis	zɛgiʃi	s	fusoʃ	fusoʃi	ʃ	pisat	pisaʧi	t
dukɛs	dukɛʃi	s	bɛdiʃ	bɛdiʃi	ʃ	zavit	zaviʧi	t
vɛdis	vɛdiʃi	s	kofuʃ	kofuʃi	ʃ	sobit	sobiʧi	t
kavus	kavuʃi	s	kikɛʃ	kikɛʃi	ʃ	dɛpot	dɛpoʧi	t
bifus	bifuʃi	s	gukoʃ	gukoʃi	ʃ	kodɛt	kodɛʧi	t
tɛbus	tɛbuʃi	s	sagoʃ	sagoʃi	ʃ	pozat	pozaʧi	t
taguz	taguʒi	z	dakuʒ	dakuʒi	ʒ			
gitaz	gitaʒi	z	gupɛʒ	gupɛʒi	ʒ			
kopaz	kopaʒi	z	vofaʒ	vofaʒi	ʒ			
tapiz	tapiʒi	z	dikaʒ	dikaʒi	ʒ			
fisuz	fisuʒi	z	fokaʒ	fokaʒi	ʒ			
sagoz	sagoʒi	z	bokiʒ	bokiʒi	ʒ			
fatɛz	fatɛʒi	z	putɛʒ	putɛʒi	ʒ			
kudɛz	kudɛʒi	z	poʃaʒ	poʃaʒi	ʒ			

Exposure: fillers								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
dɛgab	dɛgabi	b	topif	topiʃi	f	boʒup	boʒuʧi	p
kivab	kivabi	b	saʃuf	saʃuʃi	f	kuzop	kuzoʧi	p
tafɛb	tafɛbi	b	zɛkof	zɛkoʃi	f	kotap	kotapi	p
vidɛb	vidɛʒi	b	pikɛg	pikɛʒi	g	kudɛp	kudɛʧi	p
ʒadib	ʒagiʒi	b	sɛfog	sɛfogi	g	zogap	zogapi	p
fɛkud	fɛkudi	d	vatɛg	vatɛgi	g	bɛsav	bɛsaʃi	v

gavod	gavoʒi	d	zavog	zavogi	g	fizov	fizoʃi	v
gubid	gubidi	d	dapik	dapiki	k	posav	posavi	v
tuʒid	tuʒidi	d	dosuk	dosuki	k	suzɛv	suzɛvi	v
zobid	zobidi	d	puzok	puzoʧi	k	ʒɛfuv	ʒɛfuvi	v
bikaf	bikaʃi	f	vibuk	vibuʧi	k			

Generalization: test items								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
dizos	dizoʃi	s	vɛkiʃ	vɛkiʃi	ʃ	zuvit	zuviʧi	t
tɛgos	tɛgoʃi	s	tosaʃ	tosaʃi	ʃ	pibot	piboʧi	t
tazus	tazuʃi	s	kotaʃ	kotaʃi	ʃ	sapit	sapiʧi	t
bɛdas	bɛdaʃi	s	fotaʃ	fotaʃi	ʃ	ʒipɛt	ʒipɛʧi	t
vupɛs	vupɛʃi	s	kokɛʃ	kokɛʃi	ʃ	zovut	zovuʧi	t
kifɛs	kifɛʃi	s	gazuʃ	gazuʃi	ʃ	fugɛt	fugɛʧi	t
kobas	kobaʃi	s	diboʃ	diboʃi	ʃ	vaʒot	vaʒoʧi	t
gavis	gaviʃi	s	dukaʃ	dukaʃi	ʃ	ʃizut	ʃizuʧi	t
kɛtus	kɛtuʃi	s	fɛzuʃ	fɛzuʃi	ʃ	zogit	zogiʧi	t

bodas	bodaʃi	s	gafoʃ	gafoʃi	ʃ	zoʒɛt	zoʒɛʧi	t
dutaz	dutaʒi	z	sɛkoʒ	sɛkoʒi	ʒ			
fɛkiz	fɛkiʒi	z	pifaʒ	pifaʒi	ʒ			
sɛpoz	sɛpoʒi	z	zadiʒ	zadiʒi	ʒ			
govuz	govuʒi	z	dafuʒ	dafuʒi	ʒ			
ʒivuz	ʒivuʒi	z	bɛsuʒ	bɛsuʒi	ʒ			
ʃakiz	ʃakiʒi	z	vubɛʒ	vubɛʒi	ʒ			
tukɛz	tukɛʒi	z	padiʒ	padiʒi	ʒ			
bigoz	bigoʒi	z	kisɛʒ	kisɛʒi	ʒ			
tudoz	tudoʒi	z	tufiʒ	tufiʒi	ʒ			
sɛbaz	sɛbaʒi	z	pusɛʒ	pusɛʒi	ʒ			

Generalization: fillers								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
figab	figabi	b	vodag	vodagi	g	bagop	bagopi	p
dizɛb	dizɛbi	b	fɛtag	fɛtagi	g	godav	godavi	v
koʒub	koʒubi	b	fubig	fubigi	g	bogiv	bogivi	v
pavob	pavobi	b	dotag	dotagi	g	paʒɛv	paʒɛvi	v
kitub	kitubi	b	dofag	dofagi	g	fakov	fakovi	v
ʒufob	ʒufobi	b	dibɛg	dibɛgi	g	zukiv	zukivi	v
zogub	zogubi	b	zopɛg	zopɛgi	g	fupiv	fupivi	v
sazod	sazodi	d	ʒufik	ʒufiki	k	posɛv	posɛvi	v
tuvɛd	tuvɛdi	d	vizok	vizoki	k			
tɛkud	tɛkudi	d	zavuk	zavuki	k			
ʒukɛd	ʒukɛdi	d	viʒuk	viʒuki	k			
tɛbid	tɛbidi	d	tapok	tapoki	k			
gibɛd	gibɛdi	d	kutak	kutaki	k			
pizod	pizodi	d	vuzok	vuzoki	k			

kasuf	kasufi	f	sɛpak	sɛpaki	k			
ʃokuf	ʃokufi	f	bɛsap	bɛsapi	p			
sɛdaf	sɛdafi	f	gɛvip	gɛvipi	p			
tɛbif	tɛbifi	f	bafup	bafupi	p			
ʒituf	ʒitufi	f	kaʒip	kaʒipi	p			
supɛf	supɛfi	f	kɛdop	kɛdopi	p			
zɛʒif	zɛʒifi	f	gisɛp	gisɛpi	p			

b. Non-neutralizing /s/ language

Exposure: test items								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
ʒifot	ʒifoʧi	t	dɛbiʧ	dɛbiʧi	ʧ	ʒɛfos	ʒɛfoʃi	s
zuvot	zuvoʧi	t	vuzɛʧ	vuzɛʧi	ʧ	zɛʒus	zɛʒuʃi	s
pisat	pisaʧi	t	fusoʧ	fusoʧi	ʧ	zɛʧis	zɛʧiʃi	s
zavit	zaviʧi	t	bɛdiʧ	bɛdiʧi	ʧ	dukɛs	dukɛʃi	s
sobit	sobiʧi	t	ʒofuʧ	ʒofuʧi	ʧ	vɛdis	vɛdiʃi	s
dɛpot	dɛpoʧi	t	kiʒɛʧ	kiʒɛʧi	ʧ	kavus	kavuʃi	s
kodɛt	kodɛʧi	t	guʒoʧ	guʒoʧi	ʧ	biʧus	biʧuʃi	s
pozat	pozaʧi	t	sagoʧ	sagoʧi	ʧ	tɛbus	tɛbuʃi	s
taguz	taguʒi	z	ʧazuʒ	ʧazuʒi	ʒ			
gitaz	gitaʒi	z	gupɛʒ	gupɛʒi	ʒ			
ʧopaz	ʧopaʒi	z	vozaʒ	vozaʒi	ʒ			
tapiz	tapiʒi	z	dikaʒ	dikaʒi	ʒ			
fisuz	fisuʒi	z	ʧokaʒ	ʧokaʒi	ʒ			
sagoz	sagoʒi	z	bokiʒ	bokiʒi	ʒ			
fatɛz	fatɛʒi	z	putɛʒ	putɛʒi	ʒ			
kudɛz	kudɛʒi	z	poʧaʒ	poʧaʒi	ʒ			

Exposure: fillers								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
dɛgab	dɛgabi	b	saʧuf	saʧufi	f	ʧuzop	ʧuzopi	p
kivab	kivabi	b	zɛkof	zɛkofi	f	boʒup	boʒupi	p

tafɛb	tafɛbi	b	bikaf	bikafi	f	zogap	zogapi	p
ʒadib	ʒadibi	b	pikɛg	pikɛgi	g	kotap	kotapi	p
vidɛb	vidɛbi	b	vatɛg	vatɛgi	g	kudɛp	kudɛpi	p
fɛʧud	fɛʧudi	d	zavog	zavogi	g	suzɛv	suzɛvi	v
zobid	zobidi	d	sɛpog	sɛpogi	g	fizov	fizovi	v
gubid	gubidi	d	puzok	puzoki	k	posav	posavi	v
tuʒid	tuʒidi	d	dosuk	dosuki	k	ʒɛfuv	ʒɛfuvi	v
gavod	gavodi	d	vibuk	vibuki	k	bɛsav	bɛsavi	v
ʧopif	ʧopifi	f	dapik	dapiki	k			

Generalization: test items								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
zuvit	zuviʧi	t	vɛʒiʧ	vɛʒiʧi	ʧ	ʧizos	ʧizoʃi	s
pibot	piboʧi	t	tosaʧ	tosaʧi	ʧ	ʒɛgos	ʒɛgoʃi	s
sapit	sapiʧi	t	kotaʧ	kotaʧi	ʧ	ʧazus	ʧazuʃi	s
ʒipɛt	ʒipɛʧi	t	fotaʧ	fotaʧi	ʧ	bɛdas	bɛdaʃi	s

zovut	zovuʧi	t	koʒɛʧ	koʒɛʧi	ʧ	vupɛs	vupɛʃi	s
fugɛt	fugɛʧi	t	gazuʧ	gazuʧi	ʧ	ʒifɛs	ʒifɛʃi	s
vaʒot	vaʒoʧi	t	diboʧ	diboʧi	ʧ	ʒotas	ʒotaʃi	s
ʧizut	ʧizuʧi	t	duʒaʧ	duʒav	ʧ	gaʧis	gaʧiʃi	s
zogit	zogiʧi	t	fɛzuʧ	fɛzuʧi	ʧ	kɛtus	kɛtuʃi	s
zoʒɛt	zoʒɛʧi	t	gafoʧ	gafoʧi	ʧ	bodas	bodaʃi	s
duʧaz	duʧaʒi	z	sɛkoʒ	sɛkoʒi	ʒ			
fɛkiz	fɛkiʒi	z	pifaʒ	pifaʒi	ʒ			
sɛpoz	sɛpoʒi	z	zadiʒ	zadiʒi	ʒ			
govuz	govuʒi	z	daʧuʒ	daʧuʒi	ʒ			
ʒivuz	ʒivuʒi	z	bɛsuʒ	bɛsuʒi	ʒ			
ʧakiz	ʧakiʒi	z	vubɛʒ	vubɛʒi	ʒ			
tukɛz	tukɛʒi	z	paʧiʒ	paʧiʒi	ʒ			
bigoz	bigoʒi	z	kisɛʒ	kisɛʒi	ʒ			
tudoz	tudoʒi	z	tufiʒ	tufiʒi	ʒ			
sɛbaz	sɛbaʒi	z	pusɛʒ	pusɛʒi	ʒ			

Generalization: fillers								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
figab	figabi	b	vodag	vodagi	g	bagop	bagopi	p
dizɛb	dizɛbi	b	fɛʧag	fɛʧagi	g	godav	godavi	v
koʒub	koʒubi	b	fuʧig	fuʧigi	g	bogiv	bogivi	v
pavob	pavobi	b	dotag	dotagi	g	paʒɛv	paʒɛvi	v
ʧitub	ʧitubi	b	dofag	dofagi	g	fakov	fakovi	v
ʒufob	ʒufobi	b	dibɛg	dibɛgi	g	zukiv	zukivi	v
zogub	zogubi	b	zopɛg	zopɛgi	g	ʒuʧiv	ʒuʧivi	v
sazod	sazodi	d	ʒufik	ʒufiki	k	posɛv	posɛvi	v
tuvɛd	tuvɛdi	d	vizok	vizoki	k			

tɛkud	tɛkudi	d	zavuk	zavuki	k			
ʒukɛd	ʒukɛdi	d	viʒuk	viʒuki	k			
tɛbid	tɛbidi	d	tapok	tapoki	k			
gibɛd	gibɛdi	d	ʧutak	ʧutaki	k			
pizod	pizodi	d	vuzok	vuzoki	k			
kasuf	kasufi	f	sɛpak	sɛpaki	k			
ʧoʒuf	ʧoʒufi	f	bɛsap	bɛsapi	p			
sɛdaf	sɛdafi	f	gɛvip	gɛvipi	p			
ʧɛbif	ʧɛbifi	f	bafup	bafupi	p			
ʒituf	ʒitufi	f	kaʒip	kaʒipi	p			
supɛf	supɛfi	f	kɛdop	kɛdopi	p			
zɛʧif	zɛʧifi	f	gisɛp	gisɛpi	p			

c. Non-neutralizing /z/ language

Exposure: test items								
Singular	**Plural**	**final C**	**Singular**	**Plural**	**final C**	**Singular**	**Plural**	**final C**
kifot	kifoʧi	t	vokiʧ	vokiʧi	ʧ	taguz	taguʒi	z
zuvot	zuvoʧi	t	ʃɛvuʧ	ʃɛvuʧi	ʧ	gitaz	gitaʒi	z
pisat	pisaʧi	t	fosɛʧ	fosɛʧi	ʧ	kopaz	kopaʒi	z
zavit	zaviʧi	t	bidɛʧ	bidɛʧi	ʧ	tapiz	tapiʒi	z
sobit	sobiʧi	t	ʃofuʧ	ʃofuʧi	ʧ	fisuz	fisuʒi	z
dɛpot	dɛpoʧi	t	kitɛʧ	kitɛʧi	ʧ	sagoz	sagoʒi	z
kodɛt	kodɛʧi	t	gutoʧ	gutoʧi	ʧ	fatɛz	fatɛʒi	z
pozat	pozaʧi	t	sovaʧ	sovaʧi	ʧ	kudɛz	kudɛʒi	z
kɛfos	kɛfoʃi	s	dɛbiʃ	dɛbiʃi	ʃ			
zɛkus	zɛkuʃi	s	vuzɛʃ	vuzɛʃi	ʃ			
ʧɛkis	ʧɛkiʃi	s	fusoʃ	fusoʃi	ʃ			

dukɛs	dukɛʃi	s	bɛdiʃ	bɛdiʃi	ʃ			
vɛdis	vɛdiʃi	s	kofuʃ	kofuʃi	ʃ			
kavus	kavuʃi	s	kikɛʃ	kikɛʃi	ʃ			
bifus	bifuʃi	s	gukoʃ	gukoʃi	ʃ			
tɛbus	tɛbuʃi	s	sagoʃ	sagoʃi	ʃ			

Exposure: fillers

Singular	Plural	final C	Singular	Plural	final C	Singular	Plural	final C
dɛgab	dɛgabi	b	topif	topifi	f	boʃup	boʃupi	p
kivab	kivabi	b	saʃuf	saʃufi	f	kuzop	kuzopi	p
tafɛb	tafɛbi	b	zɛkof	zɛkofi	f	kotap	kotapi	p
vidɛb	vidɛbi	b	pikɛg	pikɛgi	g	kudɛp	kudɛpi	p
fadib	fadibi	b	sɛpog	sɛpogi	g	zogap	zogapi	p
fɛkud	fɛkudi	d	vatɛg	vatɛgi	g	bɛsav	bɛsavi	v
gavod	gavodi	d	zavog	zavogi	g	fizov	fizovi	v
gubid	gubidi	d	dapik	dapiki	k	posav	posavi	v
tufid	tufidi	d	dosuk	dosuki	k	suzɛv	suzɛvi	v
zobid	zobidi	d	puzok	puzoki	k	ʃɛfuv	ʃɛfuvi	v
bikaf	bikafi	f	vibuk	vibuki	k			

Generalization: test items

Singular	Plural	final C	Singular	Plural	final C	Singular	Plural	final C
zuvit	zuviʧi	t	ʧɛfiʧ	ʧɛfiʧi	ʧ	dutaz	dutaʒi	z

pibot	pibotʃi	t	vosaʧ	vosaʧi	ʧ	fɛkiz	fɛkiʒi	z
sapit	sapiʧi	t	tokaʧ	tokaʧi	ʧ	ʧɛpoz	ʧɛpoʒi	z
kipɛt	kipɛʧi	t	botaʧ	botaʧi	ʧ	govuz	govuʒi	z
zovut	zovuʧi	t	ʃobɛʧ	ʃobɛʧi	ʧ	kivuz	kivuʒi	z
fugɛt	fugɛʧi	t	dazɛʧ	dazɛʧi	ʧ	ʃakiz	ʃakiʒi	z
vakot	vakoʧi	t	sɛboʧ	sɛboʧi	ʧ	ʧukɛz	ʧukɛʒi	z
ʃizut	ʃizuʧi	t	ʃukaʧ	ʃukaʧi	ʧ	bigoz	bigoʒi	z
zogit	zogiʧi	t	kuzɛʧ	kuzɛʧi	ʧ	tudoz	tudoʒi	z
zokɛt	zokɛʧi	t	ʃofaʧ	ʃofaʧi	ʧ	sɛbaz	sɛbaʒi	z
ʃizos	ʃizoʃi	s	vɛkiʃ	vɛkiʃi	ʃ			
tɛgos	tɛgoʃi	s	tosaʃ	tosaʃi	ʃ			
tazus	tazuʃi	s	ʧotaʃ	ʧotaʃi	ʃ			
bɛdas	bɛdaʃi	s	fotaʃ	fotaʃi	ʃ			
vupɛs	vupɛʃi	s	kokɛʃ	kokɛʃi	ʃ			
kifɛs	kifɛʃi	s	gazuʃ	gazuʃi	ʃ			
kotas	kotaʃi	s	diboʃ	diboʃi	ʃ			
ʃafis	ʃafiʃi	s	dukaʃ	dukaʃi	ʃ			
kɛtus	kɛtuʃi	s	fɛzuʃ	fɛzuʃi	ʃ			
bodas	bodaʃi	s	gafoʃ	gafoʃi	ʃ			

Generalization: fillers

Singular	Plural	final C	Singular	Plural	final C	Singular	Plural	final C
figab	figabi	b	vodag	vodagi	g	bagop	bagopi	p
dizɛb	dizɛbi	b	fɛtag	fɛtagi	g	godav	godavi	v
kofub	kofubi	b	fuʧig	fuʧigi	g	bogiv	bogivi	v
pavob	pavobi	b	dotag	dotagi	g	paʒɛv	paʒɛvi	v
kitub	kitubi	b	dofag	dofagi	g	fakov	fakovi	v
tufob	tufobi	b	dibɛg	dibɛgi	g	zukiv	zukivi	v

zogub	zogubi	b	zopɛg	zopɛgi	g	ʃugiv	ʃugivi	v
sazod	sazodi	d	kufik	kufiki	k	posɛv	posɛvi	v
tuvɛd	tuvɛdi	d	vizok	vizoki	k			
tɛkud	tɛkudi	d	zavuk	zavuki	k			
tukɛd	tukɛdi	d	vituk	vituki	k			
tɛbid	tɛbidi	d	tapok	tapoki	k			
gibɛd	gibɛdi	d	kutak	kutaki	k			
pizod	pizodi	d	vuzok	vuzoki	k			
kasuf	kasufi	f	sɛpak	sɛpaki	k			
ʃokuf	ʃokufi	f	bɛsap	bɛsapi	p			
sɛdaf	sɛdafi	f	gɛvip	gɛvipi	p			
tɛbif	tɛbifi	f	bafup	bafupi	p			
kituf	kitufi	f	kaʃip	kaʃipi	p			
supɛf	supɛfi	f	kɛdop	kɛdopi	p			
zɛʧif	zɛʧifi	f	gisɛp	gisɛpi	p			

References

Bateman, N. (2007). *A crosslinguistic investigation of palatalization*. Doctoral dissertation, Univ. of California San Diego

Becker, M., & Levine, J. (2010). *Experigen–an online experiment platform*. Available (April 2013) at https://github. com/tlozoot/experigen.

Beckman, J. (1997). *Positional Faithfulness*. Retrieved from Proquest. AAI9823717.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, *167*(224), 233-277.

Bennett, R., Tang, K., & Sian, J. A. (2018). Statistical and acoustic effects on the perception of stop consonants in Kaqchikel (Mayan). *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1).

Boomershine, A., Hall, K. C., Hume, E., & Johnson, K. (2008). The impact of allophony versus contrast on speech perception. *Contrast in phonology: Theory, perception, acquisition*, *13*, 145-172.

Brown, A. (1988). Functional load and the teaching of pronunciation. *Tesol Quarterly*, *22*(4), 593-606.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 997-990.

Cychosz, M. (2017). Functional load and frequency predict consonant emergence across five languages. *UC Berkeley PhonLab Annual Report*, *13*(1).

Hall, K. C., Hume, E., Jaeger, F. T., & Wedel, A. (2016). The message shapes phonology. *Ms. University of British Columbia, University of Canterbury, University of Rochester & Arizona University*.

Harnsberger, J. D. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics*, *29*(3), 303-327.

Hockett, C. F. (1955). *A manual of phonology* (No. 11). Waverly Press.

Hualde, J. I. (2004). Quasi-phonemic contrasts in Spanish. In *WCCFL 23: Proceedings of the 23rd West Coast Conference on Formal Linguistics* (Vol. 23, p. 374). Somerville, MA: Cascadilla Press.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47-B57.

Jakobson, R. (1931). Prinzipien der historischen Phonologie. *Travaux du cercle linguistique de Prague*, 4, 246–267.

Johnson, K., & Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of Phonetics*, 38(1), 127-136.

Kaplan, A. (2011). How much homophony is normal?. *Journal of Linguistics*, *47*(3), 631-671.

Kučera, H. (1963). Entropy, redundancy, and functional load in Russian and Czech. *American contributions to the Fifth International Conference of Slavists (Sofia)*, 191-219.

Lin, S., & Demuth, K. (2015). Children's acquisition of English onset and coda /l/ : Articulatory evidence. *Journal of Speech, Language, and Hearing Research*, *58*(1), 13-27.

MacLeod, A. A., Sutton, A., Trudeau, N., & Thordardottir, E. (2011). The acquisition of consonants in Québécois French: A cross-sectional study of pre-school aged children. *International journal of speech-language pathology*, *13*(2), 93-109.

Martinet, A. (1952). Function, structure, and sound change. *Word*, *8*(1), 1-32.

Moreton, E., & Pater, J. (2012). Structure and Substance in Artificial-phonology Learning, Part I: Structure. *Language and Linguistics Compass*, *6*(11), 686-701.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, Part II: Substance. *Language and Linguistics Compass*, *6*(11), 702-718.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520-531.

New B., Pallier C., Ferrand L., Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, *L'Année Psychologique, 101*, 447-462. http://www.lexique.org

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Peperkamp, S. & Dupoux, E. (2007). Learning the mapping from surface to underlying representations in an artificial language. In J. Cole & J. Hualde (Eds.). *Laboratory phonology (Vol. 9, pp. 315-338).* Berlin: Mouton de Gruyter.

Pye, C. Ingram, D. & List, H. (1987). A Comparison of Initial Consonant Acquisition in English and Quiché. In K. Nelson & A. van Kleeck (Eds.), *Children's Language* (Vol. 6, pp. 175-190). Hillsdale, NJ: Erlbaum.

Renwick, M., & Ladd, D. (2016). Phonetic distinctiveness vs. lexical contrastiveness in non-robust phonemic contrasts. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1).

Scobbie, J. M., & Stuart-Smith, J. (2008). Quasi-phonemic contrast and the fuzzy inventory: Examples from Scottish English. *Contrast in phonology: Theory, perception, acquisition*, 87-114.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication* (1998 ed.). Urbana-Champaign: University of Illinois Press.

Silverman, D. (2009). Neutralization and anti-homophony in Korean. *Journal of Linguistics*, *46*(2), 453-482.

Silverman, D. (2012). *Neutralization.* Cambridge University Press.

Skoruppa, K., Lambrechts, A. & Peperkamp, S. (2012). The role of phonetic distance in the acquisition of phonological alternations. Lima, S., Mullin, K., Smith, B. (eds.). *Proceedings of NELS 39*, Vol. 2. CreateSpace Independent Publishing Platform. 717-729.

Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. *Ms., UCLA*.

Surendran, D., & Niyogi, P. (2003). Measuring the functional load of phonological contrasts. *arXiv preprint cs/0311036*.

Todd, S. (2012). Functional load and length-based Māori vowel contrast. Poster presented at the Annual Meeting of the New Zealand Linguistic Society. Auckland, Dec. 2012.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1-20.

Trubetzkoy, N. S. (1969). *Principles of phonology*. (C. Baltaxe, Trans.). University of California Press. (Original work published in 1939)

Vaden, K.I., Halpin, H.R., Hickok, G.S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.IPhOD.com.

Wedel, A., Kaplan A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition 128*. 179-86.

Weide, R. L. (1994). *CMU Pronouncing Dictionary*. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

White, J. (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition, 130(1),* 96-115.

White, J., & Sundara, M. (2014). Biased generalization of newly learned phonological alternations by 12-month-old infants. *Cognition, 133(1),* 85-90.

Wilson, R. H., Carnell, C. S., & Cleghorn, A. L. (2007). The Words-in-Noise (WIN) test with multitalker babble and speech-spectrum noise maskers. *Journal of the American Academy of Audiology*, *18*(6), 522-529.

Yin, H. & White, J. (2015). Neutralization avoidance and naturalness in the learning of palatalization. *Proceedings of the Annual Meeting on Phonology* 2015.

Yin, S. H., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. *Cognition*, *179*, 89-101.