GENERAL PURPOSE COGNITIVE PROCESSING CONSTRAINTS AND PHONOTACTIC PROPERTIES OF THE VOCABULARY

Padraic Monaghan[1] & Willem H. Zuidema[2]
[1]Department of Psychology, Lancaster University, Lancaster UK
[2]Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam The Netherlands

ABSTRACT

Properties of phonological systems may derive from both comprehension and production constraints. In this study, we test the extent to which general purpose constraints from sequence production are manifested in repetitions of phonemes within words. We find that near repetitions of phonemes occur less than expected by chance within the vocabularies of four studied languages: Dutch, English, French and German. This is consistent with constraints on response suppression effects in short term sequence production and with the principle of "similar place avoidance", but inconsistent with theories of consonant harmony derived from formalisation of co-articulation of phonemes in speech.

Keywords: phonotactics, production constraints, repetition, consonant harmony, co-articulation.

1. INTRODUCTION

An influential and effective approach to studies of evolutionary adaptation in structuring phonology has been to explore comprehension pressures for signal detection. For instance, Zuidema and de Boer [15] demonstrated that combinatorial phonology was an effective solution to maximizing fidelity in the acoustic form of words, and Kirby et al. [8] showed in studies of language transmission that words were likely to change from holistic forms to incorporate combinatorial structure to maximize comprehension efficiency. Similarly, Monaghan et al. [11] showed that the arrangement of form-meaning mappings in language was more easily acquired if the relation between sound and meaning was arbitrary rather than systematic. In context, an arbitrary mapping enabled the maximum information from the signal to be used in order to disambiguate the intended referent, thus enhancing the distinctiveness and signal to noise ratio for comprehension.

Each of these approaches has shown that general purpose learning mechanisms applying to comprehension result in observed patterns of phonological structure. But other properties of phonology may be the result of an accumulation of pressures from general purpose constraints on production. In this paper we address one such potential production limitation: the occurrence of repeating phonemes in the vocabulary. The distribution of repetitions within words provides insight into the communicative pressures that have resulted in the phonotactic patterns observed in extant vocabularies.

Phonological productions require a sequence of phonemes to be articulated, and as such they are prone to general purpose production constraints on sequences. One possible influential constraint is the effect of repetition on sequence encoding and/or reproduction from the memory literature. In short term memory tasks, if participants are required to recall a sequence containing a repetition then the consequence is a reduction in recall accuracy for the repeated number, particularly when it is separated by 1, 2, or 3 other numbers. This observation, known as the Ranschburg effect [3, 5], has been linked to constraints on production, as the reproduction of the sequence during recall is prone to response suppression which prohibits the same element being reproduced more than once [7]. This process is likely to result in *fewer* repetitions within words of phonemes than expected by chance.

A potentially counteractive pressure from production results from co-articulation effects, that assimilates manner or place of articulation of phonemes at points close together in the speech signal [4, 14]. This general purpose constraint on production would have the consequence that repetitions of phonemes may occur *more* than expected by chance.

What is currently lacking in the literature is a comprehensive quantitative analysis of phoneme repetitions at different positions within words in the vocabulary in order to determine whether either of these potential production constraints are affecting the phonotactic structure of the vocabulary. Given that phoneme inventories, and phonotactic constraints, co-evolve to address the joint issue of maximising perception but minimizing production effort. Thus, investigating how such constraints may relate to general purpose cognitive or articulatory limitations is key to understanding extant phonemic inventories, syllabic structure, as well as the way such phonotactic constraints can be used to support word identification, e.g., [2].

One exception is a previous study of repetition distributions in the work of MacKay [9], who showed that for subsamples of Croatian and

Hawaiian, there appeared to be a peak of repetitions for vowels one phoneme apart, and a peak for repetitions of consonants 3 phonemes apart. However, these studies were on only small samples of the corpora, and the extent to which other phonotactic constraints were driving the effects – such as the sonority hierarchy – were not possible to discern in these small-scale analyses.

Related to this is a cross-linguistic analysis of "similar place avoidance", where pairs of consonants within words are less likely to have the same place of articulation [13]. Across 30 languages, there are fewer attested forms of words containing the same place of articulation for pairs of consonants. However, the distance between phonemes that contributed to the similar place avoidance was not determined, and nor was its relation to other properties of phonemes, such as similarities in manner of articulation.

We address the issue of whether repetitions are more or less likely than chance, where we take various other constraints into account in determining a baseline, random distribution of repetitions. If the co-articulation harmony hypothesis affects phonotactic structure then we would anticipate a greater number of repetitions between phonemes in the vocabulary than chance, whereas if the Ranschburg effect affects phonotactic structure of the vocabulary, then we would expect that repetitions of phonemes close together in the word occur at a frequency less than chance.

Table 1: Properties of the vocabularies used in the analyses.

Property	Dutch	English	French	German
Number of words	117,116	53,699	62,123	79,675
Mean word length (phonemes)	9.090	6.970	6.852	8.890
Number of distinct phonemes	44	53	39	57

2. CORPUS PREPARATION

We investigated four different languages: Dutch, English, French, and German. The vocabulary lists were taken from the CELEX database [1] for Dutch, English, and German, and from Lexique 3.80 [12] for French. Only forms that were attested in the corpora used to generate frequency information were included (so for the CELEX lists, and for Lexique where frequency information was taken from the film under-titles database, frequency was > 0). Lists of words included only unique phonological forms, so homophones occurred only once in each vocabulary. Forms that comprised more than one word in the databases were also omitted, but compound forms that were listed as a single word were included. Table 1 shows the characteristics of each vocabulary.

3. REPETITIONS OF CONSONANTS AND VOWELS

3.1. Analysis

For each vocabulary, we investigated the within-word repetitions at different distances, from 1 (adjacent phonemes) to 10 (with 9 intervening phonemes) phonemes apart. Note that for larger distances, only the longer words in the vocabulary would be contributing to the counts. At each distance, the number of repetitions of phonemes was assessed. These were separated into repetitions containing consonants and those containing pairs of vowels, in order to account for different phonotactic constraints applying to vowels and consonants – i.e., vowels tend to be preceded and followed by consonants [6]. Thus, for the word "popular" (/pɒpjʊlə/), at separation distance of one, the consonant pair /pj/ would be assessed for repetitions, and at this distance this resulted in no vowel-vowel pairs. For the vowels, at this separation distance, the word contributed no repetitions. Then, for distance of 2, the pairs /pp/ and /jl/ for the consonants, and /ʊə/ for the vowels would be assessed. At this separation distance, the word contributed one repetition in the consonant analysis (/pp/). Then, repetitions at separation distance of 3 were calculated, and so on up to distance of 10 phonemes (though for the word "popular", there were no phoneme pairs assessed beyond separation distance of 6 phonemes).

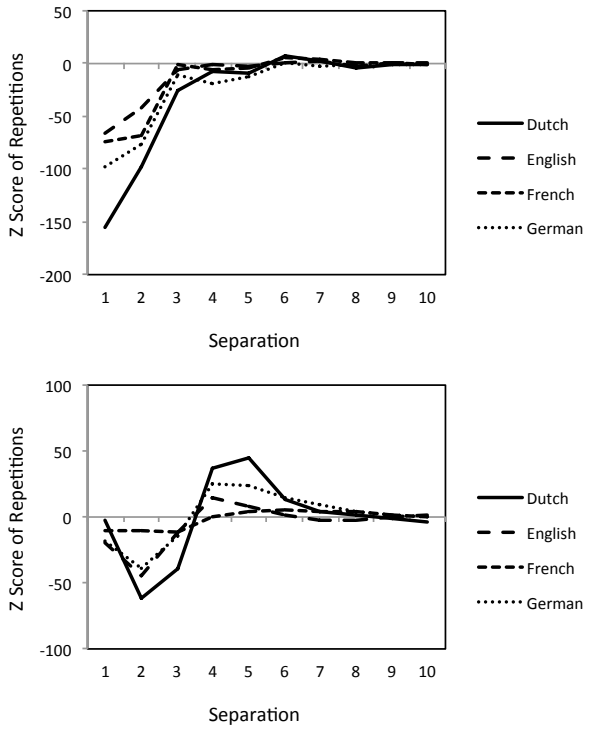
In order to determine whether the repetitions in the vocabulary occurred with frequency greater or less than expected by chance, we compared the actual number of repetitions to a random baseline distribution, where the phonemes within consonant pairs or within vowel pairs at each separation distance were randomly reassigned and the number of repetitions that occurred by chance was determined. This was repeated 10,000 times in a Monte Carlo analysis.

The resulting baseline distributions were similar to normal distributions, and so we determined the Z-score of the actual repetitions that occurred against the random distribution. Z-scores less than 0 indicate actual repetitions occur less than expected by chance, Z-scores greater than 0 indicate actual repetitions are greater than expected by chance. Z-scores > |2.81| are significantly different than chance ($p < .05$).

3.2. Results

Figure 1 shows the results of the analysis of consonant repetitions and vowel repetitions, respectively, within each vocabulary. The x-axis indicates the number of phonemes intervening between the repetition. The y-axis indicates the Z-score of the actual number of repetitions against the repetitions resulting from randomised versions of the corpus.

Figure 1: Repetitions of consonants (upper) and vowels (lower).



The results were very similar across all the languages. Not unexpectedly, there were fewer immediate repetitions than expected by chance for both vowels and consonants (separation distance 1). However, this suppression of repetition also pertained for separations up to 5 apart for the consonants, and 3 apart for vowels. For consonants separated by more than 5 other phonemes, there was variation across the languages for whether repetitions were at chance, or slightly above chance. Both Dutch and French had more repetitions than expected by chance at separation distances 5 and 6, and English and German were not significantly different than chance.

For the vowels, there was a general pattern of greater repetitions than expected by chance for separation distances 4 to 6. For longer separation distances, the distribution of repetitions converged to chance levels.

The general pattern of repetitions observed in these four languages is somewhat consistent with that of MacKay's [9] analyses of small subsets of corpora in Croatian and Hawaiian, and is in alignment with general cognitive processing constraints that drive the Ranschburg effect in short term memory tasks. Thus, across these languages, there are fewer instances of words such as "bob" or "blob", than there are words without repetitions such as "bod" or "blot".

Table 2: Mock example of calculating, separated by one other phoneme, consonant repetitions, repetitions modulated by manner, and repetitions modulated by place of articulation. Note for randomised same manner, phonemes are randomised across sets with the same manner of articulation (so only phonemes in the pairs p_g, t__p, b_b, and b_p are interchangeable, and v_v is only interchangeable with itself). For randomised same place, only phonemes in the pairs p_b and b_p are interchangeable, s_t is only interchangeable with itself, and v_v is only interchangeable with itself.

Word	Con pairs	Ran Con pairs	Same mann. pairs	Ran Same mann.	Same place pairs	Ran Same place
pop	**p_p**	**p_p**	**p_p**	p_g	**p_p**	p_b
sot	s_t	s_b			s_t	s_t
top	t_p	t_v	t_p	t_p		
bob	**b_b**	b_t	**b_b**	**b_b**	**b_b**	**b_p**
bog	b_g	b_p	b_g	b_p		
viv	**v_v**	v_g	**v_v**	**v_v**	**v_v**	**v_v**
Total Reps	3	1	3	2	3	2

4. REPETITIONS OF CONSONANTS MODULATED BY MANNER AND PLACE

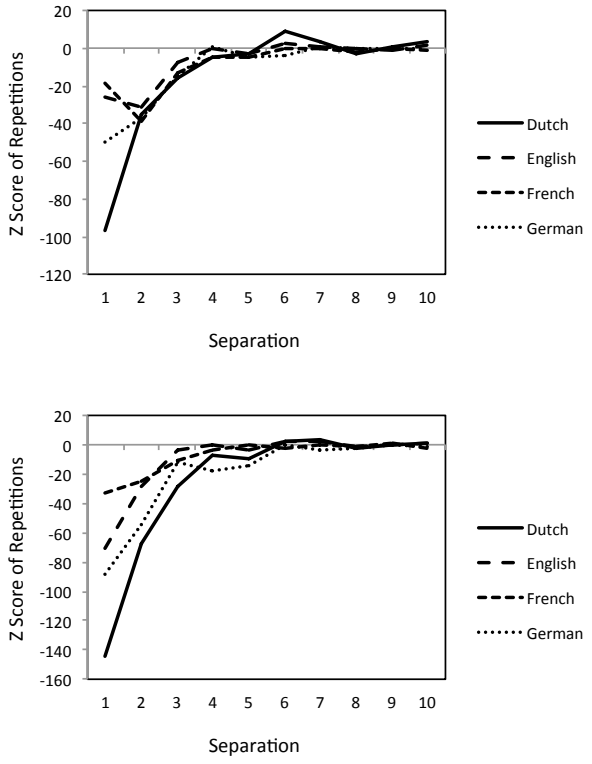
4.1. Analysis

The previous analyses of consonants take as a random baseline any two consonants, and reorder their co-occurrences. However, there are phonotactic constraints that operate over phonemes in terms of their manner of articulation that are due to syllabic structure rather than other limitations on phoneme occurrences within these structures. For instance, the sonority hierarchy permits plosive-approximant sequences in onsets but not in codas of syllables. To better respect these potential constraints that derive from the syllable structure, we repeated the analyses of consonant repetitions, but differentiated repetitions according to manner of articulation. Thus, only the phoneme pairs with similar manner of articulation were considered and random reassignments of the phonemes to these pairs occurred within manner of articulation pairs. So, for the example /pɒpjʊlə/, at separation distance 2 for the plosive manner of articulation only /p-p/ contributed to the set of plosive pairs to be

reassigned, and only /j-l/ contributed to the set of approximant pairs to be reassigned. These randomized sublists were then tested for repetitions and the results were summed and compared to the actual repetitions occurring in the vocabulary. Table 2 shows a mock example of the calculations.

A similar analysis was performed but this time considering pairs of consonants that had the same place of articulation (so for the /pɒpjʊlə/ example, /p-p/ would be entered into the set of bilabial phonemes for random reassignment but /j-l/ would not be included in a randomized set because the place of articulation differed (see Table 2 for an example). The effect of these analyses modulated by manner or by place was to inflate the repetitions of phonemes that occurred by chance in the Monte Carlo randomized analyses.

Figure 2: Repetitions of consonants modulated by manner (upper) and place (lower) of articulation.



4.2. Results

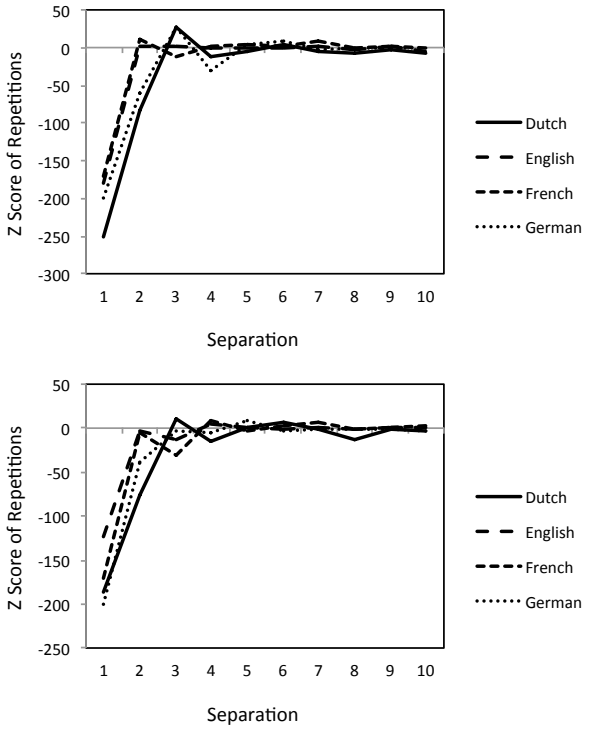
The results are shown in Figure 2 for consonant repetitions modulated by manner of articulation, and modulated by place of articulation. The effects generally reflect the previous analyses: repetitions occur less than expected by chance for consonants that occur close together in words. The avoidance of place similarity for near consonants reflected that of previous studies of the similar place avoidance principle [2, 12] but showed in addition that separation distance weakened the avoidance effect.

5. REPETITIONS OF ABSTRACT STRUCTURE

5.1. Analysis

The analyses thus far have assumed that constraints on repetition apply at the phoneme level. Thus, it is still possible for consonant harmony effects to be observed, which apply more abstractly to classes of phonemes with similar place or manner of articulation. Hence, repetitions of individual phonemes could be inhibited, but still repetitions of phonemes of the same manner of articulation could occur more than expected by chance. This would be a way in which consonant harmony effects could co-exist with reduced repetition of individual phonemes. In the final set of analyses, we assessed the extent to which repetitions of phonemes with the same manner of articulation were repeated at different separations within the vocabulary. Thus, if a plosive occurred with any other plosive that would be counted as an occurrence of a repetition. The random baseline was computed by randomly assigning phonemes to positions, but then measuring the manner of articulation of these randomly rearranged vocabularies. A similar analysis was conducted for phonemes with the same place of articulation.

Figure 3: Repetitions of phonemes with same manner (upper) and place (lower) of articulation.



5.2. Results

Figure 3 shows the results for repetitions of phonemes with the same manner of articulation and for phoneme repetitions with the same place of

articulation. The general pattern of results demonstrate that phonemes with the same manner or place of articulation tend to be inhibited at near positions in the vocabulary. However, there are some exceptions. For English, there are slightly more repetitions of phonemes with the same manner of articulation with one other phoneme separating (so "bod" is more likely than "mod"). For Dutch and German, there is a peak at distance 3 for the same manner of articulation (so "dank" would be more likely than "rank"). For place of articulation, the only repetition that occurs more than chance is for Dutch at separation distance 3. Thus, there may be some small contributions of consonant harmony effects for some of these languages, but the general effect is that there are reduced co-occurrences of phonemes with the same manner or place of articulation, again consistent with the similar place avoidance principle [12], but again that it is a graded phenomenon according to distance.

6. CONCLUSION

The starting point for these analyses was to determine whether repetitions occurred more or less than by chance, to test whether phonotactic structure was consistent with either the immediate suppression of repetitions as predicted by the Ranschburg effect, or enhancement of repetitions as predicted by co-articulation accounts of consonant harmony. The general results are more consistent with the former general purpose production constraint: close repetitions of phonemes are less likely than expected by chance within the vocabularies of the four languages we have analysed. However, this suppression effect appeared to (also) operate more abstractly in terms of suppressing repetitions of phonemes with the same manner or same place of articulation, consistent with the similar place avoidance principle [2]. Thus, there are in fact fewer consonant harmony effects than expected by chance at close distances of separation. One possible explanation for this is that co-articulation effects are actually inhibited in the vocabulary to prevent mistaken apprehension of co-articulatory effects: If the speaker produces a co-articulation then the listener can be sure that this is an error of production, therefore avoiding possible ambiguities of production [10].

These corpus analyses provide a first step to establishing the phenomena within the phonotactic structure of these languages. The next step is to confirm with experimental studies the effect of repetitions of phonemes and classes of phonemes at near and far points of repetition in the vocabulary.

7. REFERENCES

[1] Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

[2] Boll-Avetisyan, N.A.T. & Kager, R.W.J. (2014). OCP-PLACE in speech segmentation. *Language and Speech*, *57*, 394-421.

[3] Crowder, R. G. (1968). Intraserial repetition effects in immediate memory. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 446-451.

[4] Hansson, G. (2001). The phonologization of production constraints: Evidence from consonant harmony. *Chicago Linguistic Society* (Vol. 37, p. 200).

[5] Henson, R. N. (1998). Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(5), 1162.

[6] Hockema, S. A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development, 2*, 119-146.

[7] Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*(2), 117-143.

[8] Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 10681-10686.

[9] MacKay, D. W. (1970). Phoneme repetition in the structure of languages. *Language and Speech*, *13*, 199-213.

[10] McRoberts, G. W., McDonough, C., & Lakusta, L. (2009). The role of verbal repetition in the development of infant speech preferences from 4 to 14 months of age. *Infancy*, *14*(2), 162-194.

[11] Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The Arbitrariness of the Sign: Learning Advantages From the Structure of the Vocabulary. *Journal of Experimental Psychology*, 325-347.

[12] 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

[13] Pozdniakov, K., & Segerer, G. (2007). Similar Place Avoidance: A statistical universal. *Linguistic Typology*, *11,* 307–348.

[14] Rose, S., & Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, *80*, 475-531

[15] Zuidema, W. & de Boer, B. (2009), The evolution of combinatorial phonology. *Journal of Phonetics*, *37*, 125-144.