

Phonotactics as phonology:

Knowledge of a complex constraint in Dutch

René Kager (Utrecht University) and Joe Pater (University of Massachusetts, Amherst)

Abstract: The Dutch lexicon contains very few sequences of a long vowel followed by a consonant cluster, where the second member of the cluster is a non-coronal. We provide experimental evidence that Dutch speakers have implicit knowledge of this gap, which cannot be reduced to the probability of segmental sequences or to word-likeness as measured by neighborhood density. The experiment also shows that the ill-formedness of this sequence is mediated by syllable structure: it has a weaker effect on judgments when the last consonant begins a new syllable. We provide an account in terms of Hayes and Wilson's Maximum Entropy model of phonotactics, using constraints that go beyond the complexity permitted by their model of constraint induction.

1. Introduction

Phonological analysis and theorizing typically takes phonotactics as part of the data to be accounted for in terms of a phonological grammar. Ohala (1986) challenges this assumption, raising the possibility that knowledge of the shape of a language's words could be reduced to "knowledge of the lexicon plus the possession of very general cognitive abilities".¹ As an example of the requisite general cognitive abilities, he points to Greenberg and Jenkins' (1964) similarity metric, which measures the closeness of a nonce word to the existing lexicon as proportional to the number of real words obtained by substituting any set of the nonce word's segments with other segments.

Like many other models of word-likeness or probability, Greenberg and Jenkins' model of phonotactic knowledge is quite different from those constructed in phonological theory. The first difference is in the nature of the representations employed: Greenberg and Jenkins make use only of segments, while a phonological analysis of a language's phonotactics typically refers to segment-internal features as well as prosodic constituents such as syllables. While Greenberg and Jenkins and Ohala recognize the absence of features from the model as a potential flaw, purely segmental models of phonotactics continue to be

¹ We have also encountered this point of view in informal conversations with many phonologists, who speculate like Ohala that phonotactics could be reduced to knowledge of the lexicon, unmediated by phonological grammar. Thus, it would in fact be inaccurate to say that there is consensus within the phonological community that phonotactics is part of the explanatory burden of phonological grammar.

common in psycholinguistics and natural language processing. The second difference is in the nature of the computation used to evaluate a representation. As Ohala (1986) points out, the Greenberg and Jenkins model assumes no “abstracted (‘derivative’) knowledge about language-specific or language-universal sound patterns”, that is, it makes no use of phonological constraints or rules. The third and final difference is in the nature of the classification of a representation that the model returns: the Greenberg and Jenkins model yields a range of scores for representations that would typically be classified as only either “well-formed” or “ill-formed” in a phonological analysis (this two-way classification is commonly assumed in phonological theory – see e.g. Hayes 2004, Prince and Tesar 2004, though cf. Chomsky and Halle 1968:417 and others cited just below). Ohala (1986) points to experimental data gathered by Ohala and Ohala (1986) that shows that nonce word judgments are gradient in the way that the Greenberg and Jenkins model predicts.

Phonologists have only recently responded to Ohala’s (1986) challenge. The first type of response has involved the creation of models of phonotactics that provide gradient evaluation of well-formedness, but that exploit the representational vocabulary of phonological theory (e.g. Coleman and Pierrehumbert 1997, Bailey and Hahn 2001, Frisch, Pierrehumbert and Broe 2004, Anttila 2008, Coetzee 2008, Coetzee and Pater 2008, Hayes and Wilson 2008, Albright 2009). The second type of response has been to provide experimental evidence that phonotactic knowledge is encoded in terms of the constructs of phonological theory, and cannot be attributed solely to measures of word-likeness or probability calculated over segmental strings (e.g. Berent, Everett and Shimron 2001, Frisch and Zawaydeh 2001, Coetzee 2008, Albright 2009).

The present paper contributes a novel type of evidence in support of a phonological account of phonotactics. Our focus is on a phonotactic restriction whose complexity poses a challenge to both what can be taken as a baseline probabilistic model of phonotactics, as well as to some versions of the more elaborate Maximum Entropy model proposed by Hayes and Wilson (2008).

A probabilistic model of phonotactics assigns to every word some probability over the space of possible words. The simplest probabilistic model of phonotactics is an *n*-gram model. An *n*-gram model of phonotactics is based on the frequency of each segmental string of length *n* in the words of a language. These frequencies yield a probability for each of the strings (the frequency of a string divided by the summed frequencies of all strings of length *n*). The probability of any string longer than *n* can then be calculated as the product of the

probabilities of the substrings of length n . For example, a unigram model would calculate the probability of [paka] as $P([p]) \cdot P([a]) \cdot P([k]) \cdot P([a])$. A bigram model would calculate the probability of [#paka#] as $P([\#p]) \cdot P([pa]) \cdot P([ak]) \cdot P([ka]) \cdot P([a\#])$.

While appealingly simple, and broadly used in psycholinguistic research and natural language processing (see Jurafsky and Martin 2000), n -gram models are limited in terms of the types of pattern they can capture. One limit derives from the sparseness of attested strings over the space of possible strings when n gets larger than 2, making even trigrams typically unuseable as models of phonotactics (see Pierrehumbert 2003: 214 *ff.* for related discussion). A second limit derives from the inability of n -gram models to cross-classify strings, to express the non-independent restrictions imposed by constraints on various aspects of phonological representation (Martin 2007, Hayes and Wilson 2008).

To overcome these limits, Hayes and Wilson (2008) propose a Maximum Entropy model of phonotactics, which calculates probability of a word based on the weights of a set of phonological constraints (see Hayes and Wilson 2008, as well as section 4 below on the details of this calculation). There is no limit to complexity of the constraints that can be employed in a Maximum Entropy model, nor is there a requirement that they be independent. This probabilistic model of phonotactics can in principle accommodate any constraint that has been proposed in phonological theory (that is, so long as the constraint assigns to each representation a numerical score, such as the number of violations).

Hayes and Wilson (2008) also propose a method for learning the constraints, and this method does impose limits on constraint complexity. Hayes and Wilson's constraints are restrictions against sequences of segment types (classified in terms of natural classes), with an upper limit of three elements in the sequence. To express dependencies between non-adjacent segments, as needed to capture vowel harmony and stress placement regularities, Hayes and Wilson make use of projections, similar to those used in autosegmental theory. However, Hayes and Wilson's learner does not acquire constraints that refer to syllable structure or any other prosodic structure.

The Dutch phonotactic restriction that is the subject of our study targets long vowels followed by a consonant cluster ending in a non-coronal (e.g. *[me:lk]). This *V:CC_[-cor] restriction cannot be expressed in terms of bigrams, since both the sequence of a long vowel followed by a consonant (e.g. [me:l] 'flour'), as well as the consonant cluster ending in a non-coronal (e.g. [mɛ:lk] 'milk'), are well-attested. Section 2 documents this restriction based on corpus statistics.

Pierrehumbert (2003: 218) raises the issue of whether triphone constraints are internalized by native speakers: “Possibly, people may have some implicit knowledge of these facts, but it would be difficult to demonstrate that any such knowledge goes beyond lexical neighborhood effects”. Section 3 presents experimental results that show that Dutch speakers have indeed internalized the *V:CC_[-cor] restriction. In two types of judgment study (scalar rating and comparative well-formedness), subjects show a stronger dispreference for long vowels followed by a cluster than for long vowels followed by just a singleton consonant. The experimental items were controlled for bigram and trigram frequency, as well as for measures of lexical similarity, including the lexical neighborhood density measure mentioned by Pierrehumbert (which is close to the Greenberg and Jenkins model discussed above). These controls appear to rule out any model of phonotactics based purely on raw segment frequency as an account for these results.

Violations of *V:CC_[-cor] can occur both tautosyllabically, and across a syllable boundary (e.g. *[me:lkəl]). At issue is whether violations in these contexts are judged equally serious by human subjects. Our corpus study supports Kager’s (1989) observation that this sequence is also highly under-represented heterosyllabically: it gives no evidence for the constraint's sensitivity to syllable structure. Nevertheless, our experimental results show that syllable structure affects subjects’ judgments; *V:CC_[-cor] has a stronger effect when the string is contained in a single syllable.

We take these results to support the view that phonotactics do require a phonological account, since no explicit model of “knowledge of the lexicon plus...general cognitive abilities” that we know of can account for them. They also support the view that phonotactic knowledge is not categorical, since the *V:CC_[-cor] constraint does have exceptions, yet speakers still show evidence of having internalized the pattern. In section 4, we provide an account of our results in terms of Hayes and Wilson’s Maximum Entropy model of phonotactic grammar, using a constraint set that goes beyond the one that would be induced by their learner. In particular, the constraint set must distinguish between violations of *V:CC_[-cor] that occur word-finally and those that occur word-internally. In terms of purely linear constraints, this would exceed the three-element maximum imposed by Hayes and Wilson. In our account, the distinction is achieved by allowing constraints to be indexed to prosodic context. We also extend Hayes and Wilson’s model by using their grammar with an on-line learner. We do not, however, provide an alternative model of constraint induction that would generate these constraints. The case of *V:CC_[-cor] thus continues to stand as a

challenge for models of the learning of phonotactics.

2. Corpus data

We focus on a phonotactic pattern of Dutch that is known as *Moulton's generalization* (Moulton 1962). Long vowels before consonants are allowed (1a), and so are clusters ending in non-coronals (1b), yet the combination of the two substrings creates a structure that has been claimed to be ill-formed (though as we will see, there are exceptions) (Moulton 1962, Trommelen 1983).

(1)	a.	VC		b.	VCC _[-cor]		c.	*V:CC _[-cor]
		pa:l	‘pole’		palm	‘palm’		*pa:lm
		sto:m	‘steam’		stomp	‘dull’		*sto:mp
		ha:r	‘hair’		harp	‘harp’		*ha:rp
		me:l	‘flour’		mɛlk	‘milk’		*me:lk

This 'long' constraint *V:CC_[-cor] spans three segments, combining a length restriction on the first segment (a vowel) and a place restriction on the third (a consonant). A small number of exceptions occur (e.g. [twa:lf] ‘twelve’, [ʃmiŋk] ‘makeup’).² Although the pattern may appear to reflect a rime size restriction (Trommelen 1983), a complication is that this constraint seems to hold even when C₂ is in onset of a second syllable (Kager 1989) (e.g. *[me:lkəl], again with some exceptions, e.g. [kumpəl] ‘miner’).

The following are the statistics (type counts) from a dictionary of 8,305 monomorphemic stems derived from the CELEX Dutch Phonological Lemmas database (Baayen, Piepenbrock and Gulikers 1995). We start with word-final position, as in the above examples.³ The expected values (E) are calculated from the joint probability of the two levels of a variable in this set of strings. For example, E = 99 in the V:CC cell comes from the overall probability of long vowels (0.39) times the overall probability of clusters (0.19) times the total number of strings (1340). O/E is the observed number (O) divided by E. When O/E

² High tense vowels /i u y/, while phonetically short in most contexts, are distributionally long (Moulton 1962; Trommelen 1983). We have included these vowels in our corpus data as well as in our stimuli, for the reason that Moulton's generalization can be restated in terms of tenseness (rather than length), which does not affect our argument.

³ In all counts presented below, consonants C₁ in VC₁ and VC₁C₂ were limited to liquids {l r} and nasals {m n ŋ}, and consonants in C₂ of VC₁C₂ to obstruents {p b f v; k g x ɣ; t d s z} and nasals. Long vowels included {a e i o u y ø}, short vowels {ɑ ɛ ɪ ɔ œ}.

is lower than 1, the observed value is lower than expected. Since there are only two long vowels in the context of word-final clusters that end in non-coronals, O/E is quite low (O/E = 0.02). A chi-square test measures the likelihood that the overall distribution arose from chance. The distribution in Table 1 is highly unlikely to have arisen from chance (Yates chi-square = 191.87; df = 1; p < 0.0001).

	_C	_CC
V	563 (O/E = 0.85)	252 (O/E = 1.63)
V:	523 (O/E = 1.23)	2 (O/E = 0.02)

Table 1. Word-final position in monomorphemic Dutch stems, where C₂ is [-cor].

The under-representation of V:CC_[-cor] sequences is thus not merely the product of the probability of long vowels in the _CC_[-cor] context and the probability of postvocalic CC_[-cor] sequences as compared with C. It is also not the case that the gap can be explained by the low probability of clusters ending in a non-coronal (CC_[-cor]) as compared to ones ending in a coronal (CC_[cor]). This can be seen by examining the last column in the fuller contingency table in Table 2: VCC_[-cor] sequences continue to be under-represented, and V:CC_[-cor] sequences continue to be over-represented. We can further note that while long vowels are somewhat under-represented in all _CC contexts, the effect is far stronger with clusters that end in a non-coronal. The overall distribution is again highly unlikely to have arisen by chance (Chi-square = 240.38; df = 2; p < 0.0001).

	_C	_CC _[+cor]	_CC _[-cor]
V	563 (O/E = 0.80)	264 (O/E = 1.22)	252 (O/E = 1.53)
V:	523 (O/E = 1.36)	70 (O/E = 0.59)	2 (O/E = 0.02)

Table 2. Word-final position in monomorphemic Dutch stems

We now turn to the distribution of clusters which occur in prevocalic position, and hence, are heterosyllabic. Again, long vowels are highly under-represented before non-

coronal-final clusters. Compared with word-final position, the coronal/non-coronal difference is slightly smaller (that is, the degree of under-representation of long vowels before coronal-final clusters is higher in Table 3 than Table 2). The overall distribution is again highly unlikely to have arisen by chance (Chi-square = 1190; df = 2; $p < 0.0001$).

	_C	_CC _[+cor]	_CC _[-cor]
V	408 (O/E = 0.49)	485 (O/E = 1.73)	492 (O/E = 1.80)
V:	1124 (O/E = 1.61)	30 (O/E = 0.13)	10 (O/E = 0.04)

Table 3. Prevocalic position in monomorphemic Dutch stems

The bisyllabic words in our experiment are limited to monomorphemic stems ending in a schwa-liquid cluster. Table 4 presents the figures for words of this type in the Dutch lexicon, which form a subset of the cases in Table 3. The pattern is largely the same, though for some reason long vowels are preferred in the _C context in table 3, while short vowels are more common in this same context in table 4. The overall distribution is once more highly unlikely to have arisen by chance (chi-square = 23.48; df = 2; $p < 0.0001$).

	_C	_CC _[+cor]	_CC _[-cor]
V	54 (O/E = 0.86)	66 (O/E = 1.07)	71 (O/E = 1.07)
V:	14 (O/E = 2.66)	1 (O/E = 0.19)	1 (O/E = 0.18)

Table 4. Prevocalic position in Dutch stems ending in VC(C)əL

In sum, long vowels are strongly under-represented before CC_[-cor] clusters both in tautosyllabic (word-final) position and in heterosyllabic (prevocalic) position. The corpus data thus clearly support the *V:CC_[-cor] constraint, but also leave open its sensitivity to syllable structure. To investigate whether the *V:CC_[-cor] constraint is internalized by native speakers, and whether the constraint is syllable-sensitive, we ran two experiments with non-word judgments.

3. Experimental data

3.1 Experiment 1: Scalar ratings

The first experiment involved a scalar rating task, in which native speakers were presented with nonce words and indicated on a 7-point scale how word-like each stimulus sounded. Hypothesis 1 states that judgments of nonce words should be influenced by the *V:CC_[-cor] constraint, which predicts that the difference in ratings between nonce words containing a long versus a short vowel should be larger in the case of nonce words which contain a CC_[-cor] cluster than for nonce words which contain a singleton consonant. Hypothesis 2 takes into account syllabic affiliation, stating that the negative effect on well-formedness ratings of long vowels before CC_[-cor] clusters should be stronger in tautosyllabic than in heterosyllabic contexts. This predicts that there should be an interaction between the factors Cluster (cluster versus singleton) and Number of syllables (monosyllabic versus disyllabic).

3.1.1 Design

Participants. Participants were 20 native speakers Dutch, all students of Utrecht University. None reported any hearing difficulties. They were paid a small amount for participation.

Materials. A total of 240 stimuli was included, monosyllabic and disyllabic. Monosyllabic stimuli consisted of four sets of 30 nonce words of the structural types CVC, CVCC, CV:C, and CV:CC, varying in vowel length (short versus long) and the presence of a cluster (cluster versus singleton consonant). The clusters consisted of a sonorant /l r m/ in first position, and a voiceless obstruent /p f k x/ or a nasal /m/ in second position. All clusters used /lp, lk, lm, rp, rf, rm, mp, mk/ are attested in monomorphemic words; one cluster was attested only intervocalically (/mk/). Singleton consonants ending CV(:)C nonce words were /l r m/, matching postvocalic consonants of the CV(:)CC items. Disyllabic stimuli consisted of 4 sets of 30 nonce words of the structural types CVCəC, CVCCəC, CV:CəC, and CV:CCəC, all ending in schwa plus liquid. These again had /l r m/ in C₁, but a slightly wider range of choice for C₂, which also included the voiced obstruents /b v x/. Nonce words of the types CV:l(C)əl and CV:r(C)ər were avoided in order to rule out any possible influence from constraints disfavoring identical liquids. Stimuli came in short-long pairs: for each nonword containing a short vowel, there was another that was identical except having a long vowel. Moreover,

stimuli came in cluster-singleton pairs: for each nonword containing a consonant cluster, there was another that was identical except having a singleton, omitting the cluster's final consonant. Examples of nonword stimuli occur below:

CVC	CV:C	CVCC	CV:CC
bʌm	bɑ:m	bʌmk	bɑ:mk
xʌl	xo:l	xʌlm	xo:lm

CVCəC	CV:CəC	CVCCəC	CV:CCəC
dɛmər	de:mər	dɛmxər	de:mxər
jʌlər	jo:lər	jʌlbər	jo:lbər

Table 5. Examples of monosyllabic and disyllabic nonce words.

Nonce words ending in clusters contained zero or low-frequency V(:)CC_[-cor] portions. The rationale was that if speakers have internalized a constraint, then this should generalize to all V:CC_[-cor] sequences, including zero-frequent ones (such as /e:mk/). See Appendix A for a complete list of stimuli.

Throughout the stimulus set, we controlled for lexical factors and phonotactic probability, which were held constant between sets of long-short pairs, e.g. CVC versus CV:C, CVCC versus CV:CC. Two lexical factors were controlled for between the conditions: Lexical Neighborhood Density (LND; defined as the sum of logged token frequencies of a nonce word's neighbors, where neighbors were words that result from changing, inserting, or deleting a single segment) and Cohort Density (CD; defined as the sum of logged token frequencies of a nonce word's cohort members, where a nonce word's cohort was defined as all words sharing its first three segments). In addition to lexical factors, we controlled for sequential probabilities in terms of biphone and triphone transitional probabilities (TPs; TP-biphone = $p(xy) / p(x)$; TP-triphone = $p(xyz) / p(xy)$). To the 240 stimuli, 124 fillers were added, monosyllabic and disyllabic nonce words, none of which contained clusters used in the test words.

The materials were spoken by a female native speaker who was naive to the purposes of the experiment, and were recorded digitally.

Procedure. Participants were instructed to rate the word-likeness of nonce words on a 7-point scale, which was displayed on a computer screen. On the screen, extreme positions on

the scale were labeled as follows: Position 1 "Could not have been a Dutch word", Position 7 "Could very well have been a Dutch word". Stimuli were presented aurally; no orthographic representation was shown. Participants listened to stimuli through headphones at a comfortable level of loudness. They responded by a mouse click on the screen. After each response, a pause of 1000 ms. was inserted. A warning message was shown when a response was entered before the presentation of the stimulus had been completed. Stimulus order was randomized between subjects. The task was self-paced.

3.1.2 Results

The average well-formedness ratings of the nonce words are displayed in Table 6, together with the average differential scores (differences in rating between short-long pairs).

	Monosyllabic		Disyllabic	
	Singleton	Cluster	Singleton	Cluster
Short vowel	3.74	2.38	3.95	3.19
Long vowel	3.81	1.98	3.97	3.01
Δ Long-Short	+0.07	-0.40	+0.02	-0.18

Table 6. Average well-formedness ratings for nonce words in Experiment 1, with differential scores for nonce words differing in vowel length.

These averages are displayed visually in Figures 1 and 2.

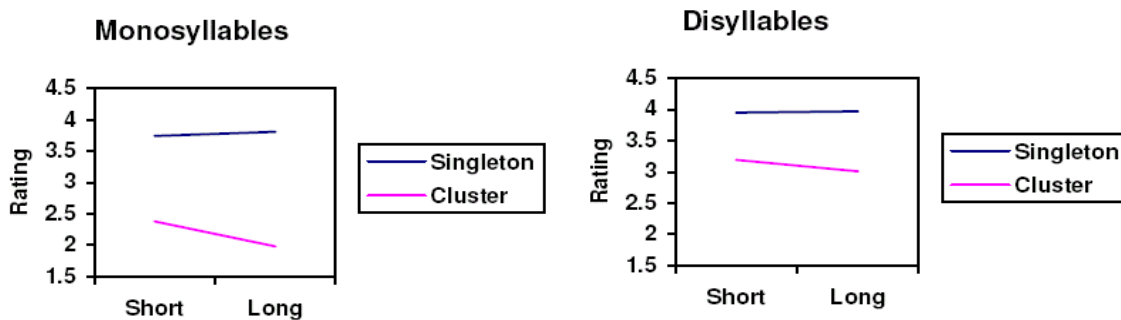


Figure 1. Average well-formedness ratings for the nonce words in Experiment 1: monosyllables and disyllables

The differential scores were analyzed by a repeated measures ANOVA, with factors Cluster (singleton *vs.* cluster) and Syllable (monosyllabic *vs.* disyllabic). Four differential scores were entered for each subject, giving a total of 80 data points: 20 subjects * 4 differential scores. A main effect of Cluster occurred ($F(1, 19) = 10.13$; $p = 0.005$) but no effect of Syllable ($F(1, 19) = 1.13$; $p = 0.302$). Posthoc paired samples t-tests on each of the monosyllabic and disyllabic sub-conditions find that the singleton *vs.* cluster difference is significant for the monosyllables ($t = 3.847$, $df = 19$, $p = 0.001$, significant at the 0.01 level with Bonferroni correction), but not for the disyllables ($t = 1.660$, $df = 19$, $p = 0.113$). The interaction of Cluster and Syllable nearly reached the 0.05 significance level ($F(1, 19) = 3.59$; $p = 0.073$).

The results confirm the prediction from Hypothesis 1 that the difference in rating between nonce words which contain a long versus a short vowel should be larger in the case of nonce words which contain a $CC_{[-cor]}$ cluster than for nonce words which contain a singleton, entailing a main effect of Cluster. The prediction from Hypothesis 2 was that there should be an interaction between the factors Cluster and Syllable. Although the results tend in the predicted direction, they fail to reach statistical significance.

3.1.3 Discussion

The results strongly support our hypothesis that Dutch native speakers have internalized a constraint $*V:CC_{[-cor]}$. Hypothesis 2, that the internalized constraint is sensitive to the syllable, received some support, in that the effect was statistically significant only in the case of the monosyllables, but the predicted interaction failed to be found. We therefore ran a second experiment that used a task that we expected would provide a more sensitive measure of subjects knowledge of word-likeness.

3.2. Experiment 2: Comparative well-formedness ratings

The second experiment involved a comparative well-formedness judgment task, in which participants were presented with pairs of nonce words and indicated which item in a pair sounded more word-like. Comparative word-likeness has been found to bring out finer-grained differences in word-likeness as compared to absolute scaling (Berent and Shimron 1997, Coetzee to appear). Hence, it was expected that comparative word-likeness ratings should bring out the syllable-dependency of the $*V:CC_{[-cor]}$ constraint more clearly. Hypotheses were identical to those of Experiment 1.

3.2.1 Design

Participants. Participants were 34 native speakers Dutch, all students of Utrecht University. None reported any hearing difficulties. They were paid a small amount for participation.

Materials. The same set of 240 monosyllabic and disyllabic nonword stimuli from Experiment 1 were used. The stimuli were paired into 120 long-short pairs, such that each stimulus pair only minimally differed in vowel length, but were otherwise identical. Examples of test pairs are:

	monosyllabic	disyllabic
CVCC - CV:CC	bɑmk - ba:mk	bɑmvər - ba:mvər
CVC - CV:C	bɑm - ba:m	bɑmər - ba:mər

Table 7. Examples of monosyllabic and disyllabic nonword test pairs used in the comparative well-formedness experiments.

The 120 test pairs were mixed with 120 filler pairs, which all consisted of one test item plus a filler item from the first experiment (e.g. [bɑmvər] - [xɔ:lər]). Each test pair and filler pair was presented in both orders, so that the total number of trials was 480. Stimulus order was randomized between participants.

Procedure. Participants were instructed to select the most word-like item of a pair of nonce words. They listened to stimuli through headphones at a comfortable level of loudness. Responses were made by a button box. The interstimulus interval (time elapsed between the members of the pairs) was 200 ms. The intertrial interval (time elapsed between the response and the next stimulus pair) was 500 ms. The response duration was 2500 ms. If no response was made at that point, the next pair was presented. Stimuli were presented with an interstimulus interval 200 msec between members of the pairs.

3.2.2 Results

For each stimulus pair, the proportion of short vowel preference was calculated by taking the ratio of the number of trials on which subjects preferred a short vowel nonce word stimulus and the total number of trials (excluding faulty trials, in which no response was made). The average short vowel preferences in Table 8 show an overall short vowel preference of 0.66 (the neutral value being 0.5). Hence, stimuli with short vowels were judged to sound more

Dutch-like than stimuli with long vowels, in all four conditions (all combinations of singleton/cluster and monosyllabic/disyllabic). Nevertheless, short vowel preferences were higher before clusters than before singletons, for monosyllables (0.732 vs. 0.585), and disyllables (0.685 vs. 0.621).

	singleton	cluster	average
monosyllabic	0.585	0.732	0.659
disyllabic	0.621	0.685	0.653
average	0.603	0.708	0.656

Table 8. Average short vowel preferences in short-long pairs differing in the number of syllables (monosyllabic versus disyllabic) and presence of a cluster (singleton versus cluster)

The difference in short vowel preference before clusters than before singletons is larger in monosyllables than disyllables, which comes out clearly in the graphical display of average short vowel preferences in Figure 2.

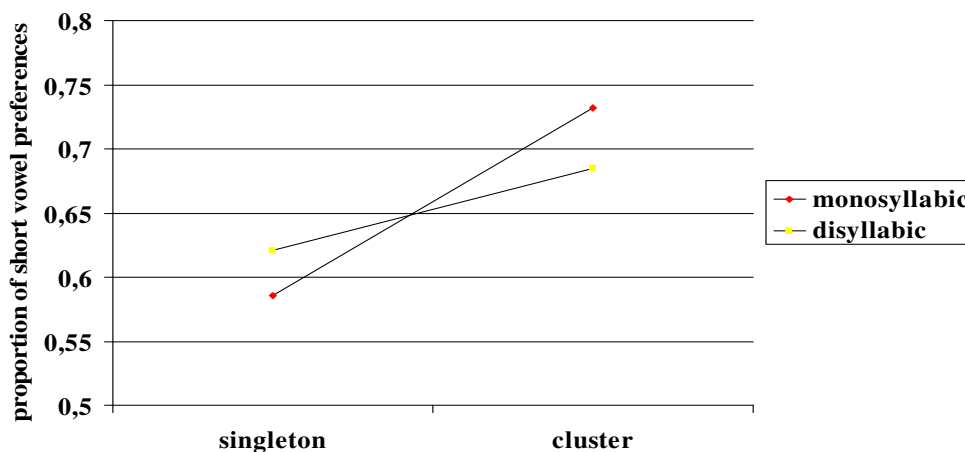


Figure 2. Average short vowel preferences in short-long pairs differing in the number of syllables (monosyllabic versus disyllabic) and presence of a cluster (singleton versus cluster)

This graphically suggests an interaction of Vowel length * Cluster. A two-way repeated-measures ANOVA on short vowel preferences with two within-subject factors, Syllable number (monosyllabic vs. disyllabic) and Cluster (singleton vs. cluster) yielded no main effect of Syllable number ($F(1, 33) = 0.021$; $p = 0.885$), but a main effect of Cluster ($F(1, 33) =$

30.035; $p = 0.000$), as well as an interaction of Syllable number and Cluster ($F(1, 33) = 15.37$; $p = 0.000$). The main effect of Cluster gives evidence for the constraint, while the interaction of Cluster and Syllable supports the constraint's syllable sensitivity. Posthoc tests find that within monosyllables the effect of cluster is significant ($t = -8.362$; $df = 33$; $p = 0.000$) while within disyllables fails to reach significance at the Bonferroni corrected significance level of 0.025 ($t = -2.092$; $df = 33$; $p = 0.044$).

Regression analysis. Although we controlled for lexical neighborhood density and sequence probability, we performed a further test of the effects of these factors by running a logistic regression model. In contrast to the ANOVA, which was limited to the test pairs differing in vowel length, this analysis was based on all responses, including responses to items from filler pairs that coincided with test items. Three phonological factors were included: Syllable number (monosyllabic vs. disyllabic), Long vowel (short vs. long), and Cluster (singleton vs. cluster). Two lexical factors were added: LND (logged lexical neighborhood density) and CD (logged cohort density), as well as transitional biphone probability (TP). Since in the comparative well-formedness design, lexical factors might affect the first and second stimulus in pairs differently, it was decided to split the three lexical factors. For example, we split LND into separate factors LND1 and LND2 (the LNDs of the first and second stimulus of the pair, respectively), and similarly for CD and TP.

The regression analysis confirms that long vowels have a strong negative effect on judgments (coef = -2.735 ; $p = 0.0000$). Presence of a cluster also has a negative effect (coef = -0.820 ; $p = 0.0034$). Number of syllables does not have an effect on its own, but it interacts with other factors. There is a strong interaction Long vowel * Cluster (coef = 1.945 ; $p = 0.0000$), such that the negative effect of long vowels is stronger in items with clusters. This is the interaction predicted from Hypothesis 1. A further interaction occurs of Long vowel * Syllable number (coef = 0.554 ; $p = 0.0003$); hence the negative effect of long vowels on well-formedness is smaller for bisyllables than for monosyllables. A marginal interaction occurs of Cluster * Syllable number (coef = 0.333 ; $p = 0.0619$), hence clusters are rated worse in monosyllables than in disyllables. Finally, a three-way interaction occurs of Long vowel * Cluster * Syllable number (coef = -0.684 ; $p = 0.0006$), predicted from Hypothesis 2. That is, the negative effect of long vowels before clusters is stronger in monosyllables than in disyllables. Of the low-level phonotactic and lexical factors, TP1 (the transitional probability of first stimuli) has a positive effect on rating of the first stimulus (i.e., the higher the TP the more "1" responses), while similarly TP2 has a negative effect (i.e., the higher the TP the

more "2" responses). However, it turns out that TP's only have an effect when other factors are in the regression model. (In a model which contains only TP1 and TP2, neither has any effect, $p > 0.15$.) LND and CD do not seem to affect the ratings at all.⁴

3.2.3 Discussion

Hypothesis 1, predicting an effect of the *V:CC_[-cor] constraint on nonword well-formedness, is again confirmed by comparative well-formedness judgments, in the form of a main effect of Cluster in the ANOVA as well as by an interaction of Long vowel * Cluster in the regression analysis. Furthermore, the constraint's effects cannot be reduced to ill-formedness of its biphone subconstituents. Although Long vowel and Cluster are both main effects, the regression analysis shows that their interaction is significant in its own right. This suggests that native speakers have internalized a triphone constraint *V:CC_[-cor]. Moreover, the regression analysis shows that effects of *V:CC_[-cor] cannot be reduced to lexical similarity as measured by LND and CD, or sequential probability, as measured by TPs.

The prediction from Hypothesis 2 was that negative effects on well-formedness of long vowels before clusters should be stronger in monosyllables, where clusters are tautosyllabic, than in disyllables, where clusters are split between syllables. This prediction was confirmed by the results. In the ANOVA, a strong interaction of Cluster and Syllable number occurred. The regression analysis supported a syllable-based interpretation of the constraint by showing a three-way interaction as predicted.

Two unexpected effects were found. First, a large negative effect of vowel length on non-word well-formedness was found, which was not anticipated. This effect cannot be entirely attributed to the low well-formedness of long vowels before clusters, since it shows up before singletons as well (see averages in Table 8). We can exclude the possibility that this reflects a general under-representation of long vowels in Dutch, since there is none. We hypothesize that the dispreference for long vowels represents an artefact of the way the stimulus pairs were constructed. Each pair contained one nonword with a short vowel and another with a long vowel. Hence, participants may have noticed this recurrent short-long vowel pattern and developed a response pattern on the basis of it, even though filler pairs had

⁴ There seems to be a small effect of CD2, but a bootstrap validation of the regression model, which is akin to a step-wise regression, deleted this factor.

been included to mask the relevant factor as much as possible.⁵ A second unexpected effect was the interaction of Long vowel and Syllable number (the negative effect of long vowels is smaller for bisyllables than for monosyllables), which may be interpreted as being due to cluster syllabification: long vowels before tautosyllabic clusters (V:CC) are judged worse than before heterosyllabic clusters (V:C.CV).

In sum, strong effects were found of vowel length and cluster (both negatively affecting non-word well-formedness) as well as an interaction of vowel length and cluster, giving evidence for the syllable-based nature of the *V:CC_[-cor] constraint.

4. A Maximum Entropy phonotactics account

Our experiments show that Dutch speakers have knowledge of the *V:CC_[-cor] restriction: their dispreference for long vowels is stronger before a cluster ending in a non-coronal than before a singleton consonant. The results also show that the force of *V:CC_[-cor] is mediated by syllable structure: the dispreference for long vowels is stronger when the cluster is fully contained in the same syllable as the vowel. In this section, we provide an account of this aspect of Dutch phonology in terms of a Maximum Entropy model of phonotactics (Hayes and Wilson 2008) that incorporates constraints referring to prosodic structure.

The Maximum Entropy model uses weighted constraints to define a probability distribution over the space of possible words. As an illustration, we consider the probability distribution defined over a set of monosyllabic words by two constraints. The constraints penalize long vowels (*V:), and consonants in coda position (*CODA), assigning a violation score of -1 for each offending structure. The scores assigned by the constraints to the types of word under consideration are shown in the rows of the table. The first word type, V:CC contains a long vowel, and thus scores -1 on *V:, as well as two coda consonants, which result in a -2 score on *CODA. The weights of the constraints are shown beneath the constraint names in the first row: *V: has a weight of 2, and *CODA has a weight of 1. The column labeled *H* shows the Harmony of each word type: the sum of the violation scores, each multiplied by the constraint's weight (see Smolensky and Legendre 2006 on the history of Harmony in linguistics). Words of the V:CC type have Harmony = 2(-1) + 1(-2) = -4. The next column *exp(H)* shows the result of raising *e* (2.72) to the power of *H*. This number is used as the basis for the calculation of the probability of each word type, shown in the final

⁵ This interpretation does not undermine the central result, as our hypotheses did not concern the presence of an effect of vowel length per se, but rather the interaction of vowel length and the presence of a cluster.

column labeled *P*.

(2) *Illustration of Maximum Entropy phonotactics*

	*V: 2	*CODA 1	<i>H</i>	<i>exp(H)</i>	<i>P</i>
V:CC	-1	-2	-4	0.02	0.01
VCC		-2	-2	0.14	0.08
V:C	-1	-1	-3	0.05	0.03
VC		-1	-1	0.37	0.22
V:	-1		-2	0.14	0.08
V			0	1.00	0.59

This example illustrates several properties of the model. First, well-formedness, defined in terms of probability, is gradient, ranging from 0.59 for a word with neither a long vowel nor coda, to 0.01 for the V:CC word type. Second, we can see how the constraint weights affect probability: because *V: has a higher weight than *CODA, words with long vowels (V:) have a lower probability (0.14) than words with codas (VC, 0.37). And finally, we can see the cumulative effect of constraint violation: *V:CC words have the lowest probability not because they violate any one constraint that the others satisfy, but because they have a greater number of constraint violations than the others.

The results that we seek to account for are summarized in Table 9. Each pair of rows corresponds to a pair of word types that were compared in our experiments. The row labels are examples of the types: [se:lk] stands for monosyllabic words with a long vowel followed by a cluster ending in a non-coronal, [se:l] stands for monosyllables with a long vowel and a final consonant, and so on. The column labeled “Δ Exp. 1” provides the difference in the mean judgment values; by subtracting the scores for the long vowel items from the short vowel ones, we get a measure of the degree of preference for the short vowels. Short vowels are preferred in the pre-[lk] environment, and the degree of that preference is higher when both consonants are in the same syllable ([se:lk] vs. [sɛlk]) than when they are not ([se:lkəl] vs. [sɛlkəl]). The “Δ Exp. 2” column shows the degree to which the mean comparative well-formedness scores diverged from chance (0.5). In all cases, short vowels were preferred over long vowels, but the degree of that preference was higher in the pre-[lk] environment, and highest of all when the consonants were tautosyllabic with the vowel.

	Δ Exp. 1	Δ Exp. 2
sɛlk > se:lk	0.40	0.23
sɛlkəl > se:lkəl	0.18	0.19
sɛl > se:l	-0.07	0.09
sɛləl > se:ləl	-0.02	0.12

Table 9. Preferences for short vowels in experimental comparisons

In terms of the Maximum Entropy model, we seek a set of constraints, and weights on those constraints, that will result in well-formedness scores being assigned to the word types that match the relative degrees of preference for short vowels across these comparisons.

We also have as a goal that the constraint weights be set on the basis of data similar to the data that a human Dutch learner would experience. As these data, we used the lexical counts presented in Section 2.⁶ The counts for each word type are repeated in Table 10. The row labels are again examples of the types: here [t] stands for coronal consonants, so that [sɛlt] stands for monosyllables with a short vowel and coronal-final cluster (see Section 2 on the exact criteria used for these categories).

⁶ For the cases in which singleton or cluster is intervocalic, we used the counts for schwa-liquid-final stems from Table 4, since these most closely matched our experimental stimuli. When we ran the simulation with the counts from Table 3, we obtained a similar, but slightly different outcome. The long-short harmony difference (see Table 11) was again the highest in the tautosyllabic pre-CC context (1.5), and the lowest in the heterosyllabic pre-C context (0.01), but the heterosyllabic pre-C and tautosyllabic pre-CC contexts are both intermediate (0.78).

<i>Word-type</i>	<i>Count</i>	<i>Proportion</i>	<i>Word-type</i>	<i>Count</i>	<i>Proportion</i>
se:lk	2	0.001	se:lkəl	1	0.001
se:lt	70	0.037	se:ltəl	1	0.001
sɛlk	252	0.134	sɛlkəl	71	0.038
sɛlt	264	0.140	sɛltəl	66	0.035
se:l	523	0.278	se:ləl	14	0.007
sɛl	563	0.299	sɛləl	54	0.029

Table 10. Learning data

The constraints penalize types of segmental sequence, as in Hayes and Wilson (2008), and much other research in phonology. We do not employ Hayes and Wilson’s constraint learning algorithm, since it does not yield prosodically conditioned constraints. Instead, we employ the full set of constraints of a particular type that apply to our case, with the expectation that these may well be learned both by humans and by a learning algorithm on the basis of the observed forms of the language.⁷ As a first pass at a set of constraints, we can consider the following “biphone”-type constraints:

- (3) *e:l Assign –1 to a long vowel followed by a consonant
 *ɛl Assign –1 to a short vowel followed by a consonant
 *lk Assign –1 to a consonant followed by a non-coronal
 *lt Assign –1 to a consonant followed by a coronal

Because constraint interaction is cumulative in the Maximum Entropy model, one might think that this constraint set could account for the *e:lk restriction as the cumulative effect of *e:l and *lk. However, the biphone constraints in a Maximum Entropy model will not suffice to account for the Dutch data, for a reason parallel to why the bigram model fails to account for

⁷ It is imaginable that constraints of this type could be projected from observations of implications that hold from single pieces of data (Burzio 2002, Boersma and Pater 2007, Adriaans & Kager 2009), but such a learning algorithm remains to be implemented. Like Hayes and Wilson (2008), we have not tailored our constraint set to yield typological predictions: with different weights, they could yield implausible outcomes, such as a language that has lower probability of short than long vowels before tautosyllabic consonants. Our results could equally well be accounted for with a typologically tailored constraint set; we adopt the present approach for ease of comparison with Hayes and Wilson (2008).

them. The problem is illustrated in (4).

(4) *The failure of biphone constraints*

	*e:l	*lk	*ɛl	<i>H</i>	<i>exp(H)</i>	<i>P</i>
	2	2	1			
se:lk	-1	-1		-4	0.02	0.03
sɛlk		-1	-1	-3	0.05	0.09
se:l	-1			-2	0.14	0.24
sɛl			-1	-1	0.37	0.64

By granting *e:l a greater weight than *ɛl, and assigning a positive weight to *lk, we succeed in making the Harmony of [se:lk] lower than that of the other word types, and thus making its probability the lowest. However, this also leads to [se:l] being at least as ill-formed relative to [sɛl] as [se:lk] is to [sɛlk]. If we take well-formedness to be a function of Harmony (see Coetzee and Pater 2008 on phonotactics), then the difference between these two pairs of word types will be equivalent (it is equal to difference between the weights of *e:l and *ɛl). If we take well-formedness to be a function of *exp(H)* (Hayes and Wilson 2008), or of *P*, then the difference between [se:l] and [sɛl] will in fact be greater, since exponentiation makes the contribution of the *e:l/*ɛl difference lower when the words also violate *lk. Either way, the outcome is not the one we want: greater ill-formedness of [se:lk] compared to [sɛlk] than of [se:l] vs. [sɛl].

We thus require a constraint that will pick out [se:lk] as especially ill-formed. To produce this constraint as one instantiation of a constraint type, we expand our constraint set to include the triphone constraints in (5):

- (5) *e:lk Assign -1 to a long vowel followed by a consonant and a non-coronal
 *ɛlk Assign -1 to a short vowel followed by a consonant and a non-coronal
 *e:lt Assign -1 to a long vowel followed by a consonant and a coronal
 *ɛlt Assign -1 to a short vowel followed by a consonant and a coronal

We also require a constraint to pick out [se:lk] as opposed to [se:lkəl], so as to make the short vowel preference stronger when the *e:lk violation is tautosyllabic. To produce the constraint

type of which this is an instance, we include domain-specific versions of all of the constraints, which apply only when the sequence is contained in a single syllable. The domain-specific version of *e:lk is shown in (6).

- (6) *e:lk_σ Assign -1 to a long vowel followed by a consonant and a non-coronal contained within a syllable

Finally, because the number of bisyllables in the learning data is low relative to the monosyllables, we include a constraint that assigns -1 to every bisyllabic word (*σσ). The presence of this constraint in our model reduces noise in the data, and allows for a better fit.

To find constraint weights, we used stochastic gradient descent, which Jäger (2006) employs for a version of Maximum Entropy grammar that operates with candidate sets, as in Optimality Theory (Prince and Smolensky 1993/2004). For Maximum Entropy phonotactics, this on-line error-driven algorithm iterates through the following steps. The learner is provided with a learning datum, or observed form, sampled from the distribution over the space of possible words. In our case, the observed form is sampled from the distribution defined by the proportions in the corpus data in Table 10. The learner then generates an expected form, by sampling from the probability distribution defined by the current grammar. The constraint weights are then updated. This is done for each constraint by subtracting the observed form's score from the expected form's score, multiplying the result by the learning rate (a positive real < 1), and adding that value to the constraint's pre-update weight. To keep the constraint weights positive, we used Exponential Maximum Entropy (Boersma and Weenink 2008), which multiplies a constraint's violation score by $exp(w)$ rather than w , where w is the constraint's weight.

The simulation we report here was performed in Praat (Boersma and Weenink 2008). The learning rate was set at 0.001, and the learner was provided with 1,000,000 pieces of data sampled from the distribution Table 10. The table in (7) shows the outcome for the word types included in our experiment. The constraints included here are all those that are violated by at least one of the word types, and have a weight $exp(w)$ greater than 0.01. The Harmony values (H) are those calculated using the unrounded weights of all of the constraints. The probability is calculated over all of the word types in the training data, and can thus be compared to those in Table 10. This comparison shows that the learner succeeded in fairly closely matching the probability of the word types in the learning data.

(7) *Results after adding triphone constraints*

	*e:lk σ 1.79	*ɛlk σ 0.31	*ɛl σ 0.07	*e:lk 3.45	*e:l 1.14	*ɛl 0.88	*lk 0.55	* $\sigma\sigma$ 2.46	<i>H</i>	<i>exp(H)</i>	<i>P</i>
se:lk	-1			-1	-1		-1		-6.93	0.001	0.001
sɛlk		-1	-1			-1	-1		-1.81	0.163	0.133
se:lkəl				-1	-1		-1	-1	-7.05	0.001	0.001
sɛlkəl			-1			-1	-1	-1	-3.41	0.033	0.027
se:l					-1				-1.14	0.320	0.262
sɛl			-1			-1			-0.95	0.386	0.316
se:ləl					-1			-1	-3.60	0.027	0.022
sɛləl						-1		-1	-3.34	0.036	0.029

Even though the learner did not have this as an explicit goal, it also succeeded in finding a grammar that lines up with our experimental results. In Table 11, we repeat the degree of preference for short vowels in the comparisons conducted in our experiments, alongside the differences between the Harmony values assigned by the final state grammar.⁸ The dispreference for long vowels is highest for the [se:lk] vs. [sɛlk] comparison, intermediate for the [se:lkəl] vs. [sɛlkəl] one, and lowest for the others.

⁸ Hayes and Wilson (2008) correlate experimental judgments of well-formedness with rescaled *exp(H)* values. To get the desired relative magnitudes of difference here, the *exp(H)* values would indeed need to be rescaled, since with the raw *exp(H)* numbers the [se:l] vs. [sɛl] difference is greater than the [se:lkəl] vs. [sɛlkəl] one. Here we can avoid rescaling by simply using Harmony rather than *exp(H)*.

	Δ Exp. 1	Δ Exp. 2	ΔH
sɛlk > se:lk	0.40	0.23	5.12
sɛlkəl > se:lkəl	0.18	0.19	3.64
sɛl > se:l	-0.07	0.09	0.19
sɛləl > se:ləl	-0.02	0.12	0.26

Table 11. Degrees of preference for short vowels in experimental results and learning results

The difference between the top two rows is due to activity of *e:lk_σ (or more precisely, the greater activity of *e:lk_σ than the sum of *ɛlk_σ and *ɛl_σ). If constraints were not relativized to syllable structure, then the only way distinguishing them would be to include constraints of a four-segment size, which also go beyond the capacity of Hayes and Wilson’s (2008) learner, and likely many other approaches to constraint induction. These results thus argue for the incorporation of syllable structure in models of phonotactic knowledge.

5. Conclusions

The results of our experiments show that native speakers have internalized a phonological constraint that refers to a sequence of three segments. The fact that this constraint cannot be construed as a combination of two biphone constraints poses problems for both *n*-gram models of phonotactics, discussed in the introduction, as well as for phonological models that induce constraints only of biphone size (see Pierrehumbert 2003⁹). A second aspect of our experimental results is that the internalized representation of this phonotactic constraint refers to the syllable. While the Hayes and Wilson (2008) model of phonotactics is in principle compatible with syllable-based phonotactics, their learning model does not induce syllabically conditioned constraints. In our learning simulation, we showed that with such constraints, a Maximum Entropy learner trained on corpus data from Dutch ends up with a grammar that matches the distinctions found in the human judgment data. This supports the general point

⁹ Pierrehumbert (2003: 218) hedges on whether a purely biphone-based theory of phonotactics is feasible, suggesting that triphone constraints might sometimes be induced. Pierrehumbert does not

made by both Pierrehumbert (2003) and Hayes and Wilson (2008) that probabilistic models, which are motivated by the gradience of phonotactic judgments, can and should operate over the sorts of representations developed in phonological theory.

References

- Adriaans, Frans & René Kager (2009). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. Submitted to *Journal of Memory & Language*.
- Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* **26**.1. 9-41.
- Anttila, Arto (2008). Gradient phonotactics and the Complexity Hypothesis. *Natural Language and Linguistic Theory* **26**, 695–729.
- Bailey, Todd & Ulrike Hahn (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* **44**, 568–591.
- Berent, Iris, Daniel L. Everett & Joseph Shimron (2001). Do phonological representations specify variables? Evidence from the Obligatory Contour Principle. *Cognitive Psychology* **42**, 1-60.
- Boersma, Paul & Joe Pater (2007). Constructing constraints from language data: The case of Canadian English diphthongs. Paper presented to the North Eastern Linguistic Society (handout available at <http://people.umass.edu/pater/boersma-pater-nels.pdf>).
- Boersma, Paul & David Weenink (2009). Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved May 1, 2009, from <http://www.praat.org/>
- Burzio, Luigi (2002). Missing players: Phonology and the past-tense debate. *Lingua*, 112:157–199.
- Chomsky, Noam & Morris Halle (1968). *The Sound Pattern of English*. New York: Harper Row.
- Coetzee, Andries (2008). Grammaticality and ungrammaticality in phonology. *Language* **84**, 218–257.
- Coetzee, Andries & Joe Pater (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* **26**, 289–337.

offer a fully explicit account of the conditions under which triphone (or even biphone) constraints are posited, so it is difficult to know whether her theory would permit *V:CC_[-cor].

- Coleman, John & Janet Pierrehumbert. 1997. 'Stochastic Phonological Grammars and Acceptability', in *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*, Association for Computational Linguistics, Somerset, NJ, pp. 49–56.
- Frisch, Stephan, Nathan Large & David Pisoni (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonce words. *Journal of Memory and Language* **42**, 481–496.
- Frisch, Stephan & Bushra Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language* **77**, 91–106.
- Frisch, Stephan, Janet Pierrehumbert & Michael Broe (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* **22**, 179–228.
- Greenberg, Joseph H. & James J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *Word* **20**, 157-177.
- Hayes, Bruce. (2004). Phonological acquisition in Optimality Theory: the early stages. In René Kager, Joe Pater and Wim Zonneveld (eds.) *Constraints in Phonological Acquisition*. Cambridge, Cambridge University Press. pp. 158-203
- Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* **39**, 379–440.
- Jäger, Gerhard. (2007). Maximum Entropy Models and Stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.), *Architectures, Rules, and Preferences. Variations on Themes by Joan W. Bresnan*. CSLI Publications, Stanford, 467-479.
- Jurafsky, Daniel & James Martin. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.
- Kager, René (1989). *A Metrical Theory of Stress and Destressing in English and Dutch*. Dordrecht: Foris.
- Martin, Andrew (2008). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. Unpublished manuscript, Laboratoire de Sciences Cognitives et Psycholinguistique.
- Moulton, William (1962). The vowels of Dutch: Phonetic and distributional classes. *Lingua* **11**, 294–312.
- Ohala, John J. (1986). Consumer's guide to evidence in phonology. *Phonology Yearbook* **3**, 2-

- Ohala, John & Manjari Ohala (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In J. Ohala and J. Jaeger (eds.) *Experimental Phonology*. Orlando, FL: Academic Press. 239–252.
- Pierrehumbert, Janet (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay and S. Jannedy (eds.) *Probability Theory in Linguistics*. Cambridge MA: The MIT Press. 177–228.
- Prince, Alan & Paul Smolensky (1993/2004). *Optimality Theory: Constraint interaction in generative grammar*. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, 2004.
- Prince, Alan & Bruce Tesar (2004). Learning phonotactic distributions. In René Kager, Joe Pater and Wim Zonneveld (eds.) *Constraints in Phonological Acquisition*. Cambridge, Cambridge University Press. pp. 245-291.
- Smolensky, Paul & Géraldine Legendre (2006). *The Harmonic Mind*. Cambridge MA: MIT Press.
- Trommelen, Mieke (1983). *The Syllable in Dutch: With special reference to diminutive formation*. Dordrecht: Foris Publications.

Appendix. Stimuli used for the nonword judgments

score short	average well-formedness judgment of the short vowel item in Experiment 1
score long	average well-formedness judgment of the long vowel item in Experiment 1
delta length	difference between the average judgments (score long – score short)
LND CVC	lexical neighborhood density of the short vowel item
LND CV:C	lexical neighborhood density of the long vowel item
CD CVC	cohort density of the short vowel item
CD CV:C	cohort density of the long vowel item
2TP short	biphone transitional probability of the short vowel item
2TP long	biphone transitional probability of the long vowel item
3TP short	triphone transitional probability of the VCC part of the short vowel item
3TP long	triphone transitional probability of the VCC part of the long vowel item

CV(:)C monosyllables

short	long	score short	score long	delta length	LND short	LND long	CD short	CD long	2TP short	2TP long
bʌm	bɑ:m	4.40	4.40	0.00	2.234	1.849	6.38	6.51	-1.202	-1.137
dɛm	dɛ:m	3.80	4.45	0.65	1.979	1.822	6.71	101.19	-1.480	-1.366
dɪl	dɪl	3.95	4.55	0.60	2.049	2.315	2.99	21.00	-1.291	-1.265
fɑr	fɑ:r	3.15	4.10	0.95	1.526	2.385	20.98	9.17	-1.301	-1.244
fɔl	fɔ:l	3.35	2.20	-1.15	2.115	2.021	31.24	6.67	-1.373	-1.427
hɔm	hʊm	4.30	3.20	-1.10	2.135	2.126	7.48	0.60	-1.316	-1.381
jɛl	jɛ:l	3.65	3.40	-0.25	2.063	2.035	0.00	0.78	-1.491	-1.523
jɪr	jɪ:r	2.60	2.10	-0.50	2.424	2.522	0.00	0.00	-1.927	-1.834
jɔr	jɔ:r	3.60	3.10	-0.50	2.373	2.587	9.13	17.44	-1.464	-1.423
kɛm	kɛ:m	4.30	3.55	-0.75	1.735	2.034	23.17	1.26	-1.429	-1.432
kɔm	kʊm	4.20	3.95	-0.25	2.021	2.017	1.15	5.42	-1.239	-1.443
lɪm	lɪm	4.05	4.00	-0.05	1.834	2.470	11.11	18.54	-1.413	-1.416
lɔm	lʊm	3.95	4.35	0.40	2.057	1.854	8.04	1.90	-1.373	-1.459
lɔr	lɔ:r	3.60	4.60	1.00	1.788	2.107	3.18	1.70	-1.363	-1.271
nɪl	nɪl	4.05	4.05	0.00	1.749	2.416	0.00	0.00	-1.402	-1.398
nɔl	nɔ:l	4.10	3.20	-0.90	2.278	1.841	2.13	0.00	-1.556	-1.471
pɔm	pʊm	3.95	4.00	0.05	1.660	1.765	3.89	2.16	-1.336	-1.458
rɪr	rɪ:r	2.25	3.65	1.40	1.823	2.485	0.00	0.00	-1.570	-1.349
sɔl	sɑ:l	3.80	3.20	-0.60	2.184	2.223	15.84	54.45	-1.375	-1.293
sɔr	sɔ:r	2.90	2.75	-0.15	1.705	2.207	18.57	25.18	-1.228	-1.194
fɔl	fɔ:l	2.55	3.35	0.80	1.886	2.050	0.00	0.00	-1.547	-1.440
tɑr	tɑ:r	4.05	4.15	0.10	1.774	2.662	26.29	22.50	-1.334	-1.246
tɪr	tɪ:r	3.90	4.90	1.00	2.173	2.351	0.00	29.92	-1.653	-1.334
xʌm	xɑ:m	3.85	4.35	0.50	2.123	1.962	8.54	5.84	-1.259	-1.242
xɛm	xɛ:m	4.45	4.30	-0.15	1.654	1.763	5.82	19.27	-1.445	-1.370
xɔl	xɔ:l	3.75	3.85	0.10	2.112	1.874	52.43	12.47	-1.267	-1.216
zʌm	zɑ:m	3.75	4.15	0.40	2.113	1.873	1.93	3.94	-1.470	-1.352
zɛl	zɛ:l	3.55	4.35	0.80	2.493	2.124	225.47	15.06	-1.275	-1.248
vɛm	vɛ:m	4.20	4.00	-0.20	2.120	1.764	0.60	16.76	-1.204	-1.198
vɔr	vɔ:r	3.75	3.75	0.00	2.299	2.614	93.27	68.61	-1.182	-1.158
mean		3.74	3.81	0.07	2.016	2.137	19.54	15.61	-1.392	-1.353
t-test		p = 0.65			p = 0.082		p = 0.65		p = 0.31	

CV(:)CC monosyllables

short	long	score short	score long	delta length	LND short	LND long	CD short	CD long	2TP short	2TP long	3TP short	3TP long
bʌmk	bɑ:mk	1.75	1.70	-0.05	2.189	1.034	6.38	6.51	-1.370	-1.318	-2.196	-2.429
dɛrx	dɛ:rx	3.40	3.30	-0.10	2.458	1.307	59.20	20.96	-1.491	-1.409	-1.615	-1.615
dɪlp	dɪlp	3.35	2.35	-1.00	0.727	1.845	2.99	21.00	-1.464	-1.442	-1.849	-2.131
fɪlk	fɪlk	2.60	2.05	-0.55	2.636	1.870	85.32	57.77	-1.489	-1.404	-2.778	-2.654
fɔlm	fɔ:lm	2.50	1.55	-0.95	1.755	2.731	31.24	6.67	-1.504	-1.547	-1.687	-1.916
hʌmk	hɑ:mk	1.30	1.55	0.25	1.777	1.561	15.32	19.19	-1.425	-1.447	-2.196	-2.429
hɔlm	hɔ:lm	2.95	2.40	-0.55	1.693	1.808	52.90	13.73	-1.451	-1.389	-1.687	-1.916
jɔelm	jʊlm	2.30	1.70	-0.60	4.043	0.670	4.04	9.50	-1.541	-1.633	-1.845	-1.589
jɔerm	jɔ:rm	2.85	1.50	-1.35	2.116	0.000	9.13	17.44	-1.509	-1.477	-1.271	-1.892
jɛrx	jɛ:rx	2.90	2.00	-0.90	2.445	0.000	0.00	2.16	-1.642	-1.598	-1.615	-1.615
jɔlm	jɔ:lm	2.10	2.15	0.05	0.778	0.000	0.60	5.43	-1.592	-1.499	-1.687	-1.916
kɛmk	kɛ:mk	2.15	1.45	-0.70	2.039	4.391	23.17	1.26	-1.552	-1.554	-2.078	-2.265
kɪrf	kɪrf	2.35	2.30	-0.05	1.044	1.632	6.39	13.24	-1.646	-1.550	-3.500	-1.957
kɪrp	kɪrp	2.70	2.60	-0.10	1.314	2.054	6.39	13.24	-1.610	-1.514	-1.445	-1.701
lɔmk	lʊmk	2.55	1.70	-0.85	0.000	0.699	8.04	1.90	-1.507	-1.576	-1.816	-1.510
lɔerm	lɔ:rm	2.85	2.65	-0.20	1.079	1.699	3.18	1.70	-1.429	-1.356	-1.271	-1.892
lɔerp	lɔ:rp	3.60	2.35	-1.25	0.602	1.699	3.18	1.70	-1.524	-1.450	-1.285	-1.684
mɪrp	mɪ:rp	2.00	2.20	0.20	0.984	2.220	2.87	28.19	-1.657	-1.489	-1.445	-1.701
nɪmk	nɪmk	2.10	1.05	-1.05	0.903	0.000	10.51	9.74	-1.644	-1.634	-2.387	-2.472
pɛrx	pɛ:rx	4.40	3.10	-1.30	2.231	1.541	217.41	49.87	-1.440	-1.419	-1.615	-1.615
pɔlm	pɔ:lm	2.40	1.55	-0.85	1.554	1.884	37.99	178.10	-1.447	-1.442	-1.687	-1.916
sʌmk	sɑ:mk	1.10	1.05	-0.05	0.301	2.660	5.46	275.11	-1.559	-1.495	-2.196	-2.429
tɔemp	tʊmp	3.35	3.20	-0.15	1.036	0.349	1.80	1.30	-1.399	-1.412	-0.971	-1.154
vʌmk	vɑ:mk	1.70	1.50	-0.20	2.241	0.477	8.94	0.48	-1.445	-1.431	-2.196	-2.429
vɛmk	vɛ:mk	1.60	1.25	-0.35	2.117	2.143	0.60	16.76	-1.371	-1.367	-2.078	-2.265
xʌmk	xɑ:mk	1.60	1.55	-0.05	1.785	0.000	8.54	5.84	-1.415	-1.402	-2.196	-2.429
xɛmk	xɛ:mk	2.05	1.95	-0.10	2.324	2.158	5.82	19.27	-1.564	-1.504	-2.078	-2.265
xɔlm	xɔ:lm	2.30	3.15	0.85	1.873	1.614	52.43	12.47	-1.419	-1.379	-1.687	-1.916
zʌmk	zɑ:mk	1.10	1.15	0.05	3.443	3.213	1.93	3.94	-1.584	-1.490	-2.196	-2.429
zɔlm	zɔ:lm	1.45	1.50	0.05	2.241	0.477	19.45	7.38	-1.509	-1.579	-1.687	-1.916
mean		2.38	1.98	-0.40	1.724	1.579	23.04	27.40	-1.507	-1.474	-1.875	-2.002
t-test		p = 0.037			p = 0.30		p = 0.74		p = 0.13		p = 0.26	

CV(:)CəL disyllables

short	long	score short	score long	delta length	LND short	LND long	CD short	CD long	2TP short	2TP long
bʊmər	bɑ:mər	4.40	4.10	-0.30	2.046	2.056	6.38	6.51	-1.032	-0.988
dɛmər	dɛ:mər	4.45	4.40	-0.05	1.823	1.609	6.71	101.19	-1.167	-1.091
fʊrəl	fɑ:rəl	3.30	3.35	0.05	2.061	1.914	20.98	9.17	-1.229	-1.191
fɪrəl	fɪ:rəl	3.00	3.60	0.60	0.000	1.173	9.73	6.86	-1.493	-1.269
hɛmər	hɛ:mər	4.45	4.55	0.10	1.418	1.685	20.53	56.94	-1.089	-1.057
jœrəl	jy:rəl	3.45	2.75	-0.70	0.000	1.415	9.13	17.44	-1.338	-1.310
jɔlər	jɔ:lər	4.65	4.30	-0.35	1.412	1.190	0.60	5.43	-1.232	-1.155
kœmər	kumər	4.35	3.95	-0.40	2.381	2.118	1.15	5.42	-1.056	-1.192
kɛmər	kɛ:mər	4.40	4.40	0.00	1.714	1.822	23.17	1.26	-1.133	-1.135
kɪrəl	kɪ:rəl	3.85	3.70	-0.15	2.076	1.942	6.39	13.24	-1.393	-1.313
lœmər	lumər	4.65	4.25	-0.40	1.571	1.412	8.04	1.90	-1.146	-1.203
lœrəl	ly:rəl	3.15	3.05	-0.10	1.517	1.699	3.18	1.70	-1.271	-1.210
lɪmər	limər	4.35	4.30	-0.05	1.508	1.707	11.11	18.54	-1.122	-1.124
nœrəl	nɑ:rəl	3.65	3.70	0.05	0.977	2.297	20.62	30.44	-1.291	-1.175
nɪlər	nɪlər	4.10	3.90	-0.20	1.610	1.628	0.00	0.00	-1.228	-1.225
nɔrəl	nɔ:rəl	3.15	3.25	0.10	2.445	1.658	49.96	77.20	-1.287	-1.222
pœmər	pumər	4.10	4.40	0.30	1.991	1.944	3.89	2.16	-1.121	-1.202
pʊlər	pɑ:lər	4.05	4.35	0.30	1.286	1.677	36.30	47.61	-1.036	-1.025
pɪlər	pɪlər	4.15	4.05	-0.10	1.386	1.400	14.59	18.63	-1.151	-1.104
pɔlər	pɔ:lər	4.70	4.75	0.05	1.346	1.721	37.99	178.10	-1.111	-1.107
rɛlər	rɛ:lər	4.65	5.00	0.35	1.802	1.094	7.10	57.74	-1.065	-1.062
sɪrəl	sɪ:rəl	2.40	3.25	0.85	2.263	1.643	59.36	44.90	-1.418	-1.147
sɔlər	sɔ:lər	3.15	3.10	-0.05	1.132	1.372	52.40	37.16	-1.124	-1.110
sœrəl	sɔ:rəl	3.20	3.35	0.15	2.178	1.415	18.57	25.18	-1.180	-1.158
tœmər	tumər	4.15	4.00	-0.15	1.237	1.481	1.80	1.30	-1.189	-1.200
vʊmər	vɑ:mər	3.95	3.90	-0.05	2.087	2.829	8.94	0.48	-1.094	-1.082
wɛlər	wɛ:lər	4.45	4.65	0.20	1.509	1.516	253.73	14.82	-0.977	-1.039
wɛmər	wɛ:mər	4.25	4.80	0.55	1.440	1.546	0.60	16.76	-1.033	-1.029
xʊmər	xɑ:mər	4.25	4.35	0.10	1.776	1.993	8.54	5.84	-1.069	-1.058
xɔlər	xɔ:lər	3.65	3.65	0.00	1.060	1.022	52.43	12.47	-1.088	-1.054
mean		3.95	3.97	0.02	1.568	1.666	25.13	27.21	-1.172	-1.141
t-test		p = 0.88			p = 0.44		p = 0.85		p = 0.27	

CV(:)CCəL disyllables

short	long	score short	score long	delta length	LND short	LND long	CD short	CD long	2TP short	2TP long	3TP short	3TP long
bamvər	ba:mvər	2.25	2.55	0.30	0.000	0.000	6.38	6.51	-1.175	-1.138	-2.196	-2.384
dəmxər	de:mxər	2.65	2.85	0.20	0.000	0.000	6.71	101.19	-1.268	-1.202	-1.953	-2.323
dilkər	dilkər	4.00	3.40	-0.60	2.468	1.954	2.99	21.00	-1.223	-1.207	-2.778	-2.654
dilpər	dilpər	4.30	3.35	-0.95	0.000	2.564	2.99	21.00	-1.275	-1.260	-1.849	-2.131
fəlbər	fai:lbər	2.90	3.30	0.40	0.000	0.000	10.12	14.53	-1.270	-1.242	-1.809	-1.771
fərxər	fə:rxəl	3.00	2.45	-0.55	0.000	0.000	20.98	9.17	-1.215	-1.183	-1.707	-2.104
fərvəl	fə:rvəl	2.65	2.75	0.10	0.000	0.000	20.98	9.17	-1.250	-1.218	-2.546	-2.444
hömvər	humvər	2.85	2.30	-0.55	0.000	0.000	7.48	0.60	-1.240	-1.277	-2.255	-1.949
jörməl	jy:rməl	2.95	2.80	-0.15	0.477	0.000	9.13	17.44	-1.376	-1.353	-1.271	-1.892
jölbər	jö:lbər	3.05	2.70	-0.35	0.000	0.000	0.60	5.43	-1.344	-1.277	-1.837	-1.854
jölmər	jö:lmər	2.90	3.30	0.40	0.000	0.000	0.60	5.43	-1.378	-1.311	-1.687	-1.916
köembər	kumbər	4.30	3.90	-0.40	0.000	0.000	1.15	5.42	-1.093	-1.210	-1.627	-1.074
köemkər	kumkər	3.40	2.35	-1.05	0.000	0.000	1.15	5.42	-1.244	-1.361	-1.816	-1.510
kəmxr	ke:mxər	3.25	2.70	-0.55	0.699	0.000	23.17	1.26	-1.238	-1.240	-1.953	-2.323
kirfəl	kirfəl	2.60	2.20	-0.40	0.000	0.000	6.39	13.24	-1.488	-1.420	-3.500	-1.957
kirxəl	kirxəl	3.60	3.90	0.30	0.000	0.000	6.39	13.24	-1.313	-1.244	-2.290	-2.144
kirpəl	kirpəl	3.75	2.90	-0.85	0.000	0.000	6.39	13.24	-1.437	-1.369	-1.445	-1.701
lörbəl	ly:rbəl	2.80	2.60	-0.20	0.000	0.000	3.18	1.70	-1.295	-1.243	-1.345	-1.674
limxər	limxər	2.80	2.25	-0.55	0.000	0.000	11.11	18.54	-1.229	-1.231	-3.500	-2.648
nörbəl	no:rbəl	3.60	4.00	0.40	0.000	1.869	49.96	77.20	-1.309	-1.253	-2.086	-1.324
pəlbər	pe:lbər	3.50	3.65	0.15	0.000	0.000	22.12	12.44	-1.160	-1.195	-1.783	-1.878
rəlxər	re:lxər	3.70	2.70	-1.00	0.000	0.000	7.10	57.74	-1.160	-1.156	-1.646	-1.735
sörvəl	so:rvəl	3.20	3.10	-0.10	0.000	0.000	18.57	25.18	-1.208	-1.189	-1.981	-1.771
tömpər	tumpər	4.00	4.00	0.00	1.158	0.778	1.80	1.30	-1.229	-1.238	-0.971	-1.154
wəmxər	we:mxər	3.15	3.05	-0.10	0.000	0.000	0.60	16.76	-1.153	-1.149	-1.953	-2.323
wəmkər	we:mkər	3.15	3.25	0.10	2.690	1.114	0.60	16.76	-1.224	-1.221	-2.078	-2.265
xəmkər	xa:mkər	3.55	3.50	-0.05	0.000	0.000	8.54	5.84	-1.255	-1.246	-2.196	-2.429
xəmvər	xa:mvər	2.85	2.60	-0.25	0.000	0.602	8.54	5.84	-1.207	-1.198	-2.196	-2.384
xəlbər	xö:lbər	2.30	2.55	0.25	0.000	0.000	52.43	12.47	-1.220	-1.191	-1.837	-1.854
xəlmər	xö:lmər	2.75	3.30	0.55	0.000	0.000	52.43	12.47	-1.254	-1.225	-1.687	-1.916
mean		3.19	3.01	-0.18	0.250	0.296	12.35	17.58	-1.258	-1.242	-1.993	-1.983
t-test		p = 0.20			p = 0.79		p = 0.29		p = 0.41		p = 0.94	