THE GRADIENT PHONOTACTICS OF ENGLISH CVC SYLLABLES

Olga Dmitrieva

ABSTRACT

This paper focuses on the factors affecting gradient well-formedness of English CVC syllables. The study examines a gradient OCP-place effect in English CVC words and syllables extracted from two electronic dictionaries, CMU and CELEX. It was found that a gradient restriction on the co-occurrence of homorganic consonants operates in all CVC syllables of English. Previously this was established only for monomorphemic monosyllabic words (Berkley 1994). The distance effect, reported by Berkley (1994), was only partially confirmed in this study. Syllables containing long vowels or diphthongs in many cases exhibited a weaker OCP effect than those containing short vowels, but this difference was only marginally significant. The effect of vowel height was also examined as a possible factor affecting the strength of the OCP. The results showed a non-significant trend in the predicted direction. Another factor affecting the well-formedness of the CVC syllables is the prominence alignment between syllable stress, vowel height, and consonant place. Stressed syllables combine more often with more sonorous low vowels. Unstressed syllables prefer less sonorous high and reduced vowels and coronal onsets and codas. It was established that the syllables violating these requirements were under-represented in the lexicon.

Following Coetzee and Pater (2008) we fit a linear regression model to the data and propose an OT approach based on partially ordered grammar (Anttila 1998). Both the statistical model and the OT grammar feature constraints capturing OCP related restrictions and prominence alignment related restrictions on the well-formedness of English CVC syllables. The proposed OT account provides a reasonably close fit to the quantitative patterns observed in English phonotactics.

1. Introduction

This paper examines the factors affecting gradient well-formedness of English CVC syllables: the gradient OCP-place (Obligatory Contour Principle) and the prominence alignment between syllable stress, vowel height, and consonant place.

The OCP-place is a gradient prohibition against homorganic consonants in the onset and the coda of the CVC syllable. For example, a word *gag* in English violates OCP-place because both onset and coda consonants are dorsal. A word *gap*, on the other hand, does not violate OCP-place because it has a dorsal consonant in the onset and a labial consonant in the coda. Obligatory Contour Principle was first proposed by Leben (1973) and given its name by Goldsmith (1976). Originally it was intended as a prohibition against two adjacent identical tones, or a requirement that melodies must be alternating. Later the OCP was extended to the segmental level and used to motivate a prohibition against identical or homorganic consonants separated by a vowel in a number of languages, including Arabic (Greenberg 1950, McCarthy 1986a, 1988, Yip 1988, Frisch *et al.* 2004, Pater and Coetzee 2005, Coetzee and Pater 2008), Muna (Coetzee and Pater 2008), English (Berkley 1994) and others. These studies showed that in some cases the prohibition is categorical while in others it is gradient. Berkley (1994) examined the behavior of homorganic consonants in monomorphemic monosyllabic words of English. She found that words, which begin and end with homorganic consonants, were underrepresented in the English lexicon. Their number was smaller than would be expected given that phonemes can combine freely. In addition, labial and dorsal co-occurrences in English were less tolerated than coronal co-occurrences.

Berkley (1994) also found that consonant pairs separated by a single vowel, such as in *pip* or *king*, were subject to a stronger OCP effect than pairs separated by a diphthong or by a vowel-consonant combination, for example, *pipe* or *cling*. Any additional intervening segments weakened the OCP effect further, as in words like *trite*, where a consonant is followed by a diphthong. When the number of intervening segments was four or more the OCP was no longer

significant. I will refer to this finding in the rest of the paper as *distance effect*. The following table summarizes Berkley's results in terms of Observed/Expected frequency ratio.

Table 1: Ratio of observed values to expected values for homorganic consonant pairs (Berkley 1994).

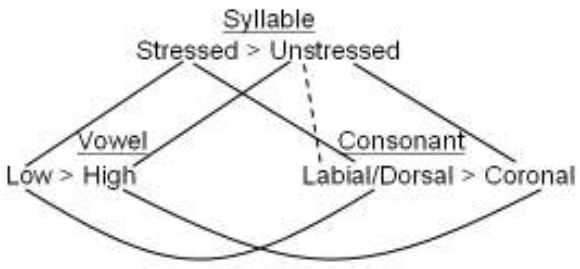
	Number of intervening segments			
	one	two	three	four or more
Labial	**0.40**	**0.52**	**0.57**	**n/a**
Dorsal	**0.42**	**0.73**	**0.56**	**n/a**
Coronal	**0.86**	**0.96**	**0.94**	**0.91**

The present paper partially replicates Berkley's findings on a different set of data and extends her treatment of the OCP from monosyllabic monomorphemic words to all CVC syllables of English. We hypothesize that CVC syllables that violate the OCP-place, that is, where C1 and C2 share the same place of articulation, are under-represented in English. In addition, following Berkley I argue that less distance between the consonants strengthens the OCP effect. Therefore, syllables with short vowels that violate the OCP-place are more under-represented than syllables with long vowels or diphthongs that violate OCP-place. Similarly, syllables with high or reduced vowels that violate OCP-place are expected to be more under-represented than syllables with low or mid vowels that violate OCP-place. This extension of the distance effect hypothesis is based on the assumption that high and reduced vowels are acoustically weaker than low and mid vowels. They are intrinsically shorter, less loud and therefore less sonorous than low and mid vowels. As a result they introduce less distance between the consonants.

Prominence alignment between syllable stress, vowel height, and consonant place was an additional factor tested against the gradient well-formedness of English CVC syllables. The definition of prominence alignment used here follows Prince and Smolensky (1993/2004: 139-151), which was based on syllable peak and margin and the sonority heirarchy scale. They showed that syllable peaks, as more prominent positions than syllable margins, prefer to contain more sonorous segments. Syllable margins, conversely, are less prominent and prefer less sonorous segments.

Based on the previous literature my analysis proposes binary prominence distinctions in each domain: stressed syllables are more prominent than unstressed syllables; low and mid vowels are more prominent than high and reduced vowels; labial or dorsal consonants are more prominent than coronal consonants. Unlike the OCP, which prohibits co-occurrence, prominence alignment requires that the most prominent element in each domain combines with the most prominent elements in other domains (see Figure 1).

Figure 1: Prominence alignment between syllable stress, vowel height, and consonant place.



Mutual attractions (solid lines in Figure 1) are the preferred alignments of the elements in English CVC syllables, and I hypothesize that any combination violating these assumptions (dashed lines in Figure 1) should be avoided and underrepresented in the lexicon. In particular, stressed syllables with high vowels, or unstressed syllables with low vowels, are expected to be under-represented compared to stressed syllables with low/mid vowels, or unstressed syllables with high/reduced vowels. Similarly, stressed syllables with coronal consonants and unstressed syllables with labial or dorsal consonants are expected to be under-represented compared to their more harmonic counterparts – stressed syllables with labial and dorsal consonants or unstressed syllables with coronal consonants. Naturally, low vowels in combination with coronal consonants or high vowels in combination with dorsal and labial consonants are predicted to be dispreferred in the lexicon in favor of low vowels combined with labials and dorsals or high vowels combined with coronals.

Unlike the OCP effect, the prominence alignment has never been suggested to play a role in the phonotactics of English syllables. This study provides a description of prominence alignment in English and uncovers evidence that, together with the OCP effect, it contributes crucially to the well-formedness of the CVC syllables.

2. Methods

Material

Two online dictionaries provided material for the representation of the English lexicon: the Carnegie Mellon University Pronunciation Dictionary, CMU (Weide 1998) and the CELEX lexical database of English (Baayen et al. 1995). There are two major differences between the dictionaries. First, CMU is a North American English dictionary and CELEX is a British English dictionary, so each can be viewed as roughly representative of the lexicons of these two general dialects of English and their variants. A second crucial difference between corpora is that CMU includes words and all their inflectional variants, while the CELEX lemma lexicon contains only derivational variants of the wordforms. Running the same analysis on these two dictionaries allows for tests of robustness across different dialects and whether different morphological factors can alter the results.

The two dictionaries encode the data in different formats and require different kinds of pre-processing. The CMU dictionary is a machine-readable pronunciation dictionary for North American English that contains 127,069 entries and their phonemic transcriptions. The version of CMU used in this study was further annotated by Michael Speriosu (Stanford University). Syllable boundary information, stress tags, and syllable weight tags were added in addition to the original information. In the resulting data file each entry had the following format:

(1) Word Pronunciation /# Syllable boundaries #/ S:stress W:weight
e.g. CITY S IH1 T IY0 /# [S '1 IH] [T '0 IY] #/ S:PU W:LH

The majority of the lexemes in the dictionary had multiple entries corresponding to different syllabifications, such that a word like *city* was represented in three entries as [ci][ty], [cit][y], and [ci[t]y], with an ambisyllabic [t] in the last entry. Multiple syllabifications were removed for the current project: only the first syllabification was preserved as the most standard. Each syllable in CMU was marked for stress: primary stress, secondary stress, and unstressed.

The CELEX lexical database contains a variety of tags for each lexical entry, including orthographic (written form), phonological (pronunciation), morphological (part of speech),

frequency, stress, and syllable type (in CV notation). In this study we were interested in phonological transcription, syllable boundaries, and stress markers. The data file had the following format:

(2) Number\word C=part of speech F=frequency S=stress P=phonemic transcription B=syllable boundaries T=type of pronunciation (primary or secondary) N=number of pronunciations
e.g. 7643\city C=1 F=4619 S='sI-tI P=[sI][tI] B=[CV][CV] T=P N=1

For the analysis of the monosyllabic words all CVC words were extracted from the dictionaries: 2,984 CVC words in CELEX and 4,958 CVC words in CMU. For the syllable analysis all CVC syllables were extracted from the dictionaries: 83,798 CVC syllables in CMU and 25,888 CVC syllables in CELEX. Each syllable was preceded by a stress marker: P(rimary), or U(nstressed). Syllables with secondary stress were excluded from the analysis, along with syllables containing diphthongs and nasalized vowels.

(3) CMU	CELEX	
U[S IH V]	U[sɪv]	as in *aggressive*
S[T IY K]	S[ti:k]	as in *antique*

Procedure

All onsets and codas were classified by place as coronal [t, d, θ, ð, s, z, ʃ, ʒ, ʧ, ʤ, r, l, n], dorsal [k, g, ŋ, j], or labial [p, b, f, v, m, w]. All vowels were classified by height as *high*, which included all high vowels and schwa [ə, ɪ, i:, ʊ, u:], or *low*, which included all non-high, low and mid, vowels [æ, ɛ, ʌ, ɑ, ɔ:, a:]. They were also classified by length as short vowels [ɪ, ʊ, æ, ɛ, ʌ, ɑ], long vowels [i:, u:, ɔ:, a:], or diphthongs [eɪ, aɪ, oɪ, aʊ, əʊ, ɪə, ɛə, ʊə]. All the syllables were classified as stressed or unstressed. This classification represents the independent factors considered in this study. Each factor has several levels. For example, consonant place factor: three levels (coronal, dorsal, labial), vowel height factor: two levels (high and low), and stress factor: two levels (stressed and unstressed). Various combinations of these factors produce a range of CVC syllable types available in English.

The degree of the representation of each syllable type in the lexicon, and the degree of its well-formedness, was quantified in terms of the Observed/Expected frequency[1] ratio (O/E ratio; Frisch *et al.* 2004). Observed frequency was the actual frequency of a given syllable type occurring in the database. Expected frequency represented a frequency of a given syllable type calculated on the assumption that levels of each factor can combine freely. For example, the probability of the syllable type dorsal-vowel-dorsal is calculated by multiplying the probability of the dorsal onset by the probability of the dorsal coda, based on the assumption that they are independent events. Expected frequency, then, is calculated by multiplying this probability by the total of the corpus.

(4) *P(dorsal-V-dorsal) = P(onset=dorsal) * P(coda=dorsal)*
*E(dorsal-V-dorsal) = P(dorsal-V-dorsal) * Total*

[1] Token frequency. For example, each occurrence of the syllable [gVg] in the dictionary was counted towards the overall frequency of the syllable type dorsal-vowel-dorsal.

In order to calculate expected frequency of the syllable type with an additional factor an extra term was added to the equation. For example, expected probability of *stressed* syllables with dorsal onsets and dorsal codas is calculated by multiplying the probability of the dorsal onset by the probability of dorsal coda by the probability of the syllable being stressed.

(5) *P(dorsal-stressed V-dorsal) = P(onset=dorsal) * P(coda=dorsal) * P(syllable=stressed)*
*E(dorsal-stressed V-dorsal) = P(dorsal-Vstressed-dorsal) * Total*

An O/E ratio greater than 1.00 signifies that there are more such syllable types observed than expected, i.e., the syllable type is over-represented. An O/E ratio smaller than 1.00 means there are fewer observed syllables of this kind than expected, i.e. the syllable type is under-represented.

In addition to the O/E ratio two statistical techniques were used to evaluate the effect size: a Chi-square test and a multiple regression analysis. While the Chi-square test can determine whether there is a relationship between any two factors, the multiple regression evaluates the contribution of each factor in a single unified analysis.

3. Results

Monosyllabic words

To examine the distribution of the consonants, onsets and codas of the monosyllabic CVC words in CMU and CELEX are divided into three places of articulation and an O/E value for each combination of the three types of consonants was calculated. The question is whether words with onsets and codas sharing place of articulation are under-represented (have O/E values below 1.00) in comparison to words with onsets and codas with different places of articulation.

As expected, the combinations of onsets and codas with the same place of articulation in CVC words were under-represented compared to the words with heterorganic onsets and codas, in both corpora. Tables 2 and 3 present the O/E values for consonant co-occurrences divided by short vowels in CMU and CELEX.

Table 2: Distribution of consonants in monosyllabic CVC words of CMU (separated by a single short vowel), e.g, *pip* /pɪp/

		Coda			
Onset		Labial	Dorsal	Coronals	
	Labial	87.00 152.64 **0.57**	141.00 146.15 0.96	562.00 491.21 1.14	Observed Expected O/E
	Dorsal	118.00 111.10 1.06	88.00 106.38 **0.83**	369.00 357.53 1.03	Observed Expected O/E
	Coronal	313.00 254.27 1.23	267.00 243.47 1.10	736.00 818.27 **0.90**	Observed Expected O/E

$\chi^2(4, N = 2681) = 66.694, p < .001$, Cramer's V = 0.112

Table 3: Distribution of consonants in monosyllabic words of CELEX (separated by a single short vowel), e.g., *pip* /plp/

		Coda			
		Labial	Dorsal	Coronal	
Onset	Labial	67.00 101.19 **0.66**	92.00 101.53 0.91	266.00 222.28 1.20	Observed Expected O/E
	Dorsal	60.00 53.33 1.13	38.00 53.51 **0.71**	126.00 117.15 1.08	Observed Expected O/E
	Coronal	168.00 140.48 1.20	166.00 140.95 1.18	256.00 308.57 **0.83**	Observed Expected O/E

$\chi^2(4, N = 1239) = 45.848, p < .001$, Cramer's V = 0.136

The results suggest that the prohibition against homogenous consonants is especially strong for labials. It is the weakest for coronals, with dorsal in an intermediate position.

The strength of the OCP was also tested in connection with the amount of the material intervening between the consonants, i.e., the distance effect (Berkeley 1994). While Berkley compared the OCP effect for consonant pairs separated by one, two, or more segments, she did not distinguish between long and short vowels, collapsing them under a one-segment category. She also joined diphthongs and vowel-consonant combinations under a two-segment category. Here, I divide vowels into short and long and consider diphthongs as a separate category, calculating an O/E value for each combination of onset, coda, and vowel length. This allows us to see whether changing only the amount of vocalic material, without additional consonants, may modify the strength of the OCP.

Tables 4 and 5 present the O/E values for OCP violations according to consonant place and vowel length.

Table 4: O/E values for consonant co-occurrences with short vowels, long vowels, and diphthongs (CMU).

	Amount of intervening material		
	Short vowel	Long vowel	Diphthong
Labial	**0.57**	**0.72**	**0.60**
Dorsal	**0.83**	**1.27**	**0.55**
Coronal	**0.90**	**0.98**	**0.94**

Table 5: O/E values for consonant co-occurrences with short vowels, long vowels, and diphthongs (CELEX).

	Amount of intervening material		
	Short vowel	Long vowel	Diphthong
Labial	**0.66**	**0.63**	**0.82**
Dorsal	**0.71**	**1.05**	**0.65**
Coronal	**0.83**	**0.91**	**1.00**

Our results show the expected trend towards higher O/E values with long vowels and diphthongs. However, this trend is has exceptions: in CMU, for example, the diphthongs have consistently lower O/E values than long vowels, or, in the case of dorsals, even short vowels.

A linear regression analysis was performed to evaluate the robustness of the quantitative trends. The model used here is adopted from Coetzee and Pater (2008). It uses a transformed observed frequency (natural logarithm) as a dependent variable in the regression and a transformed expected frequency (natural logarithm) as one of the independent variables. It is

important to include expected frequency as one of the predictors to control for the natural asymmetries in the dataset (for example, the fact that coronals are the most frequent consonants).

The regression also introduces a number of categorical predictor variables to test the strength of the OCP effect and the distance effect. To determine whether having homorganic consonants significantly reduces the frequency of a CVC form, the test includes a binary variable which has a value 1.00 in all such cases and a zero otherwise. The test also includes a number of contrast variables to determine whether the strength of the OCP restriction is significantly different for the three places of articulation. A similar set of contrast variables is applied to testing the difference between the OCP violations in the presence of short vowels, long vowels, or diphthongs, to test whether the difference in the amount of vocalic material between the consonants significantly affects the strength of the OCP. All possible combinations of the onset place, coda place, and vowel length were used as the cases in the analysis.

The results for CMU showed that the OCP factor had a significant effect: CVC words with OCP violations were significantly less frequent than words without OCP violations ($B = -0.126$, $t = -5.265$, $p < 0.001$). At the same time, the frequency of the words violating the OCP-labial was significantly different from the frequency of the words violating the OCP-dorsal and the OCP-coronal ($B = -0.069$, $t = -3.311$, $p < 0.01$). The frequency of the words with homorganic consonants separated by short or long vowels was significantly different form the frequency of the words with homorganic consonants separated by diphthongs ($B = -0.068$, $t = -3.307$, $p < 0.01$). The R value for this regression was 0.994 ($F(4, 26) = 462.027$, $p < 0.001$).

In CELEX, the OCP factor had a significant effect as well, in the expected direction ($B = -0.137$, $t = -5.819$, $p < 0.001$). The frequency of the words with OCP-labial violations was significantly different from the frequency of the words with OCP-coronal violations ($B = -0.055$, $t = -2.18$, $p < 0.05$). The difference between OCP-labial and OCP-dorsal plus OCP-coronal was near-significant ($B = -0.27$, $t = 1.954$, $p = 0.063$). In contrast to the results from CMU, none of the factors testing the distance effect reached significance. The R value for this regression was 0.991 ($F(3, 26) = 431.812$, $p < 0.001$).

The following analysis was performed to test whether the differences in vowel height can produce the effect similar to Berkley's (1994) distance effect. That is, whether intrinsically longer low and mid vowels can weaken the OCP effect in comparison with high vowels. Tables 6 and 7 present the O/E values for the consonant co-occurrences in the onsets and codas of the CVC words, divided by vowel height. In both dictionaries there was a trend towards a higher O/E values for low vowels than for high vowels in words with OCP violations. This suggests that low vowels may indeed weaken the OCP effect.

Table 6: Distribution of consonants in CVC words of CMU (separated by HIGH or LOW vowels), e.g., *pip* /pɪp/, *pop* /pop/

		Coda						
		Dorsal		Labial		Coronal		
Onset		HIGH	LOW	HIGH	LOW	HIGH	LOW	
	Dorsal	13.00 22.10 **0.59**	75.00 84.28 **0.89**	17.00 23.08 0.74	101.00 88.02 1.15	58.00 74.28 0.78	311.00 283.25 1.10	Observed Expected O/E
	Labial	35.00 30.36 1.15	106.00 115.79 0.92	17.00 31.71 **0.54**	70.00 120.93 **0.58**	140.00 102.05 1.37	422.00 389.16 1.08	Observed Expected O/E
	Coronal	60.00 50.58 1.19	207.00 192.88 1.07	58.00 52.83 1.10	255.00 201.44 1.27	159.00 170.00 **0.94**	577.00 648.26 **0.89**	Observed Expected O/E

Table 7: Distribution of consonants in CVC words of CELEX (separated by HIGH or LOW vowels), e.g., *pip* /pɪp/, *pop* /pop/

		Coda						
		Dorsal		Labial		Coronal		
Onset		HIGH	LOW	HIGH	LOW	HIGH	LOW	
	Dorsal	9.00 13.78 0.65	29.00 39.74 0.73	10.00 13.73 0.73	50.00 39.60 1.26	27.00 30.16 0.90	99.00 86.99 1.14	Observed Expected O/E
	Labial	19.00 26.14 0.73	73.00 75.39 0.97	16.00 26.05 0.61	51.00 75.14 0.68	95.00 57.23 1.66	171.00 165.05 1.04	Observed Expected O/E
	Coronal	43.00 36.29 1.18	123.00 104.66 1.18	43.00 36.17 1.19	125.00 104.31 1.20	57.00 79.45 0.72	199.00 229.12 0.87	Observed Expected O/E

The regression analysis for the CMU data showed that the OCP factor significantly reduced the frequency of the CVC words ((B = -0.170, t = -4.661, $p < 0.001$). The frequency of words with the OCP-labial violations was near-significantly different from the frequency of those with the OCP-dorsal and the OCP-coronal violations (B = -0.046, t = -2.146, p = 0.050). However, the factor testing the effect of vowel height on the magnitude of the OCP did not reach significance. The R value for this regression was 0.990 ($F(3, 17) = 237.767$, $p < 0.001$).

For the CELEX data only the OCP factor reached significance (B = -0.181, t = -4.582, $p < 0.001$). There were no significant differences between the strength of the OCP violations for labial, dorsal, and coronal co-occurrences. The effect of the vowel height on the strength of the OCP was not significant either. The R value for this regression was 0.983 ($F(2, 17) = 213.901$, $p < 0.001$).

Discussion

This analysis replicated and extended Berkley's (1994) analysis of the OCP in the monosyllabic words, revealed a similar pattern. Words with combinations of the homorganic consonants in onsets and codas are under-represented in the lexicon. This under-representation is statistically significant in both dictionaries used in the study. In addition, there is evidence that the OCP prohibition in English CVC words is especially strong against labial-labial co-occurrences: the O/E values for the words violating OCP-labial were consistently the lowest and this difference emerged as statistically significant in the regression analysis. The same trend was present in Berkley's results, as well as in other investigations of the OCP effect (McCarthy 1988, Coetzee and Pater 2008, Berkley 1994). Dorsal co-occurrences pattern together with coronal co-occurrences, particularly in the CMU results, which may be due to the fact that dorsals, as coronals, participate in morphological structures (e.g. [ɪŋ] as in *walking*).

At the same time, the numbers suggest a higher tolerance towards the OCP violations than in Berkley's study. The dataset used in this study included morphologically complex words while Berkley's data contained only monomorphemic words. It is possible that the OCP effect is weaker across morphemes than within morphemes, explaining the higher O/E values in the results of this study.

In terms of the distance effects the results were inconclusive. While there appeared to be a trend towards a weaker OCP effect with long vowels and diphthongs, it was inconsistent and not significant statistically. This contradicts Berkley's (1994) findings. However, in Berkley's analysis short vowels were not distinguished from long vowels and diphthongs were collapsed with vowel-consonant combinations. As a result, Berkley's study was often comparing the strength of the OCP effect with a single intervening vowel to the strength of the OCP effect with

one or more intervening consonants. Together the results of this study and Berkley's findings suggest that only additional intervening consonants can effectively weaken the OCP.

Finally, the results showed that vowel height moderated the magnitude of the OCP effect. Comparison of the O/E values revealed a quantitative trend in the expected direction although this difference was not significant in the statistical analysis.

CVC syllables

Based on the evidence that the OCP influences the phonotactics of CVC words, the hypothesis can be extended to all CVC syllables. Tables 8 and 9 present O/E values for onset-coda co-occurrences in CVC syllables of CMU and CELEX. They reveal a robust pattern of under-representation for the syllables with homorganic onsets and codas. As in CVC words, labial co-occurrences have the lowest O/E values, coronal co-occurrences have the highest O/E values, and dorsals are in an intermediate position.

Table 8: O/E values for onset-coda co-occurrences in CVC syllables of CMU (separated by a single long or short vowel), e.g., *pip* /pɪp/, *peep* /pi:p/

		Coda			
		Labial	Dorsal	Coronal	
Onset	Labial	544.00 1740.35 **0.31**	2850.00 3380.98 0.84	20079.00 3500.57 1.09	Observed Expected O/E
	Dorsal	1587.00 972.08 1.63	965.00 1888.47 **0.51**	10559.00 10250.45 1.03	Observed Expected O/E
	Coronal	4082.00 3500.57 1.17	8255.00 6800.56 1.21	34877.00 36912.88 **0.94**	Observed Expected O/E

$\chi^2(4, N = 83{,}798) = 2438.136$, $p < .001$, Cramer's V = 0.121

Table 9: O/E values for onset-coda co-occurrences in CVC syllables of CELEX (separated by a single long or short vowel), e.g., *pip* /plp/, *peep* /pi:p/

	Coda				
		Labial	Dorsal	Coronal	
Onset	Labial	457.00 4449.56 **0.42**	1521.00 1464.97 1.04	5027.00 4449.56 1.13	Observed Expected O/E
	Dorsal	667.00 632.65 1.05	509.00 849.91 **0.60**	2888.00 2581.44 1.12	Observed Expected O/E
	Coronal	2906.00 2306.88 1.26	3384.00 3099.12 1.09	8529.00 9413.00 **0.91**	Observed Expected O/E

$\chi^2(4, N = 25{,}888) = 884.889$, $p < .001$, Cramer's V = 0.131

Tables 10 and 11 summarize the results of a test for the distance effect. There is a trend for the higher O/E values with long vowels and diphthongs, and this trend appears to be even more consistent than for the CVC words.

Table 10: O/E values for consonant co-occurrences with short vowels, long vowels, and diphthongs (CMU).

	Amount of intervening material		
	Short vowel	Long vowel	Diphthong
Labial	**0.29**	**0.36**	**0.69**
Dorsal	**0.48**	**0.82**	**0.76**
Coronal	**0.94**	**0.98**	**0.99**

Table 11: O/E values for consonant co-occurrences with short vowels, long vowels, and diphthongs (CELEX).

Amount of intervening material			
	Short vowel	Long vowel	Diphthong
Labial	**0.32**	**0.76**	**0.48**
Dorsal	**0.62**	**0.36**	**1.00**
Coronal	**0.90**	**1.00**	**0.98**

The regression analysis showed that for the CMU data the OCP factor had a significant effect on the frequency of the syllables ($R = 0.990$, $F(3, 17) = 237.767$, $p < 0.001$): syllables violating the OCP were significantly less frequent than syllables without the OCP violations ($B = -0.222$, $t = -5.039$, $p < 0.001$). The frequency of the words with the OCP-labial and the OCP-dorsal violations was significantly different from the frequency of the words with the OCP-coronal violations ($B = -0.126$, $t = -4.527$, $p < 0.001$). The factors testing the distance effect did not reach significance.

The regression analysis for the CELEX data again showed that OCP is a significant factor for syllables ($B = -0.224$, $t = -4.736$, $p < 0.001$). In partial contrast with CMU, the frequency of syllables with violations of OCP-labial was significantly different from the frequency of syllables with the OCP-coronal and the OCP-dorsal violations ($B = -0.06$, $t = -2.168$, $p < 0.05$). Also unlike CMU, there was a near-significant difference between the frequency of the syllables with the homorganic consonants separated by short vowels and the frequency of the syllables with the homorganic consonants separated by diphthongs ($B = -089$, $t = -1.791$, $p = 0.087$).

The effect of stress on the OCP in syllables and the stress-consonant alignment

Another factor relevant for the phonotactics of the CVC syllables, but not monosyllabic words, is stress. To compare the OCP effect in stressed and unstressed syllables we calculated O/E values shown in Tables 12 and 13. Syllables with secondary stress were excluded from this analysis.

Table 12: O/E values for onset-coda co-occurrences in stressed and unstressed CVC syllables of CMU.

		Coda						
		Labial		Dorsal		Coronal		
		Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed	
Onset	Labial	474.00	77.00	1234.00	1629.00	8586.00	11596.00	Observed
		603.20	1137.15	1171.83	2209.15	6360.61	11991.07	Expected
		0.79	**0.07**	1.05	0.74	1.35	0.97	O/E
	Dorsal	984.00	579.00	476.00	428.00	4469.00	5718.00	Observed
		336.92	635.16	654.53	1233.93	3552.76	6697.69	Expected
		2.92	0.91	**0.73**	**0.35**	1.26	0.85	O/E
	Coronal	2164.00	1977.00	2293.00	6070.00	8577.00	27015.00	Observed
		1213.28	2287.29	2357.04	4443.51	12793.83	24119.04	Expected
		1.78	0.86	0.97	1.37	**0.67**	**1.12**	O/E

Table 13: O/E values for onset-coda co-occurrences in stressed and unstressed CVC syllables of CELEX.

		Coda						
		Labial		Dorsal		Coronal		
Onset		Stressed	Unstressed	Stressed	Unstressed	Stressed	Unstressed	
	Labial	336.00 435.76 **0.77**	121.00 654.71 **0.18**	734.00 585.41 1.25	787.00 879.56 0.89	1155.00 921.84 1.25	2585.00 2671.49 0.97	Observed Expected O/E
	Dorsal	408.00 252.81 1.61	259.00 379.84 0.68	223.00 339.63 **0.66**	286.00 510.28 **0.56**	1340.00 1031.56 1.30	1548.00 1549.88 1.00	Observed Expected O/E
	Coronal	1155.00 921.84 1.25	1751.00 1385.04 1.26	1180.00 1238.43 0.95	2204.00 1860.69 1.18	2527.00 3761.49 **0.67**	6002.00 5651.51 **1.06**	Observed Expected O/E

The OCP appears to be weaker in stressed than in unstressed syllables for labial and dorsal co-occurrences. Interestingly, the opposite is true for coronal co-occurrences: the OCP is weaker in unstressed syllables than in stressed syllables. At the same time, all syllables that begin with a labial or a dorsal are under-represented in unstressed syllables compared to stressed syllables. In contrast, the syllables that begin with a coronal are under-represented in stressed syllables as compared to unstressed. This is the evidence for the prominence alignment between stress and consonant place.

To test the strength of the connection between the stress factor and the consonant place factor we performed a chi-square for the relationship between onset place and stress. It was significant in both CMU and CELEX datasets:

(6) CMU: $\chi^2(2, N = 83{,}798) = 3069.088$, $p < .001$, Cramer's V = 0.191.
CELEX: $\chi^2(2, N = 25{,}888) = 741.658$, $p < .001$, Cramer's V = 0.169.

Cramer's V value is higher for CMU than for CELEX, meaning that the association between the two variables is stronger in CMU. This alignment is illustrated in Figure 2.

Figure 2: Distribution of onset consonants across stressed and unstressed syllables of CMU

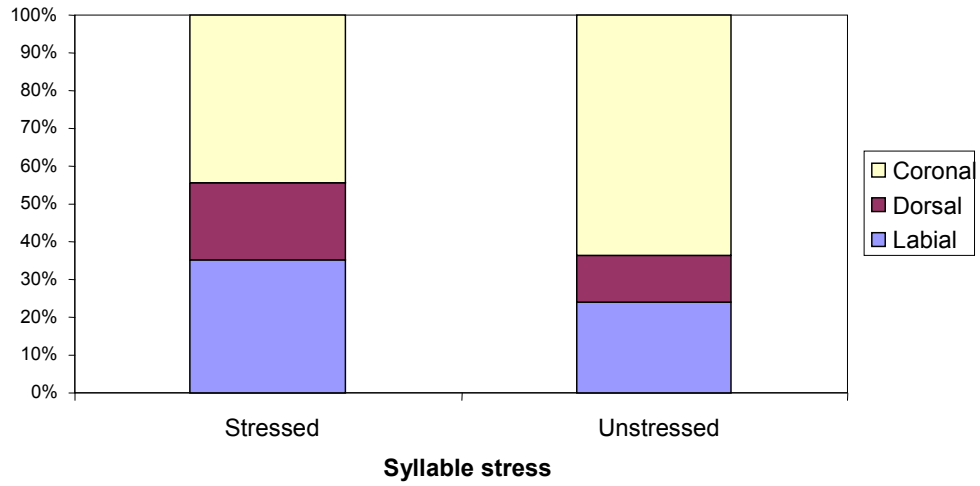


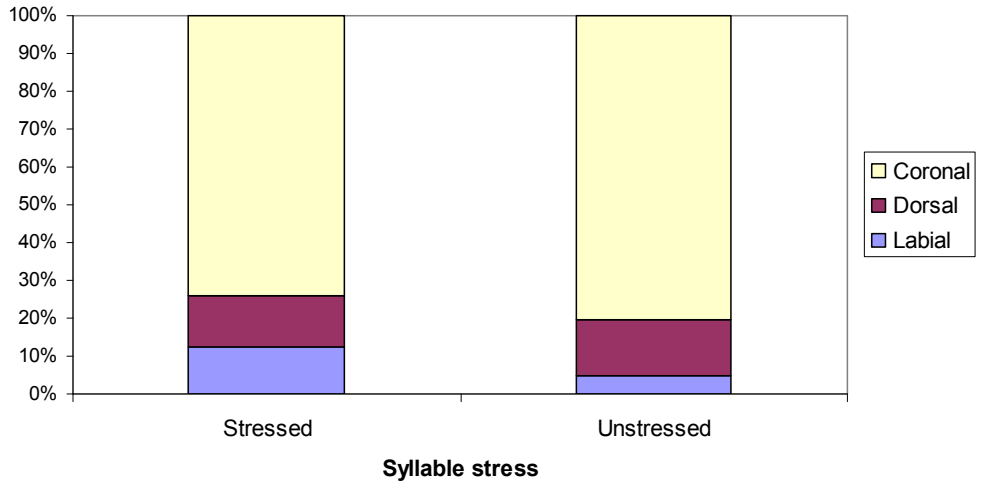
Figure 2 shows that dorsal and labial onsets are more prominent in stressed syllables than in unstressed syllables, while the frequency of syllables with coronal onsets is greater in unstressed syllables compared to stressed syllables.

A chi-square for the association between coda place and stress was also performed:

(7) CMU: $\chi^2(2, N = 83{,}798) = 1586.509$, $p < .001$, Cramer's V = 0.138.
CELEX: $\chi^2(2, N = 25{,}888) = 104.09$, $p < .001$, Cramer's V = 0.063.

Cramer's V again demonstrates that this association is stronger in CMU than in CELEX. In addition, these results show that the association between coda place and stress is weaker than between onset place and stress. Figure 3 illustrates the connection between the two variables in CMU. It shows that there is a higher proportion of coronals in unstressed syllable than in stressed syllables, although the difference is less pronounced than for the onsets. Labials are more represented in stressed syllables than in unstressed syllables, but dorsals seem to pattern together with coronals.

Figure 3: Distribution of coda consonants across stressed and unstressed syllables of CMU



The effect of vowel height on the OCP in syllables and the vowel-consonant alignment

The CVC syllables were separated into two groups according to type of vowel (low/mid or high/reduced). The distribution of the consonants in onsets and codas of the syllables showed that the O/E values were higher for labial and dorsal co-occurrences with low vowels than with high vowels. The tendency in the opposite direction was found for coronals: the O/E values are higher for syllables with high or reduced vowels than for syllables with low or mid vowels. At the same time all syllables with dorsal or labial onsets were over-represented with low vowels in comparison to high vowels. Syllables with coronal onsets, on the other hand, were over-represented with high vowels in comparison to low vowels. The results are presented in Tables 14 and 15.

Table 14: Distribution of consonants in the CVC syllables of CMU (separated by high or low vowels).

		Coda						
		Dorsal		Labial		Coronal		
Onset		HIGH	LOW	HIGH	LOW	HIGH	LOW	
	Dorsal	144.00 1260.77 **0.11**	407.00 479.58 **0.85**	575.00 704.21 **0.82**	1055.00 931.69 **1.13**	13674.00 13294.55 **1.03**	6508.00 5057.12 **1.29**	Observed Expected O/E
	Labial	575.00 704.21 **0.82**	988.00 267.87 **3.69**	524.00 2617.35 **0.20**	380.00 520.40 **0.73**	6425.00 7425.76 **0.87**	3762.00 2824.69 **1.33**	Observed Expected O/E
	Coronal	2342.00 2535.93 **0.92**	1799.00 964.64 **1.86**	6287.00 4926.54 **1.28**	2076.00 1874.01 **1.11**	29300.00 26740.89 **1.10**	6292.00 10171.99 **0.62**	Observed Expected O/E

Table 15: Distribution of consonants in the CVC syllables of CELEX (separated by high or low vowels).

		Coda						
		Dorsal		Labial		Coronal		
Onset		HIGH	LOW	HIGH	LOW	HIGH	LOW	
	Dorsal	122.00 673.79 **0.18**	335.00 416.68 **0.80**	722.00 905.19 **0.80**	799.00 559.78 **1.43**	2748.00 2749.35 **1.00**	2279.00 1700.21 **1.34**	Observed Expected O/E
	Labial	226.00 390.91 **0.58**	441.00 241.74 **1.82**	276.00 959.52 **0.29**	233.00 324.76 **0.72**	1372.00 1595.05 **0.86**	1516.00 986.39 **1.54**	Observed Expected O/E
	Coronal	1958.00 1425.41 **1.37**	948.00 881.48 **1.08**	2276.00 1914.92 **1.19**	1108.00 1184.20 **0.94**	6296.00 5816.22 **1.08**	2233.00 3596.78 **0.62**	Observed Expected O/E

The Chi-squares for the relationship between onset place and vowel height were significant for both CMU and CELEX:

(8) CMU: $\chi^2(2, N = 83{,}798) = 2584.998$, $p < .001$, Cramer's V = 0.176.
CELEX: $\chi^2(2, N = 25{,}888) = 1290.02$, $p < .001$, Cramer's V = 0.223.

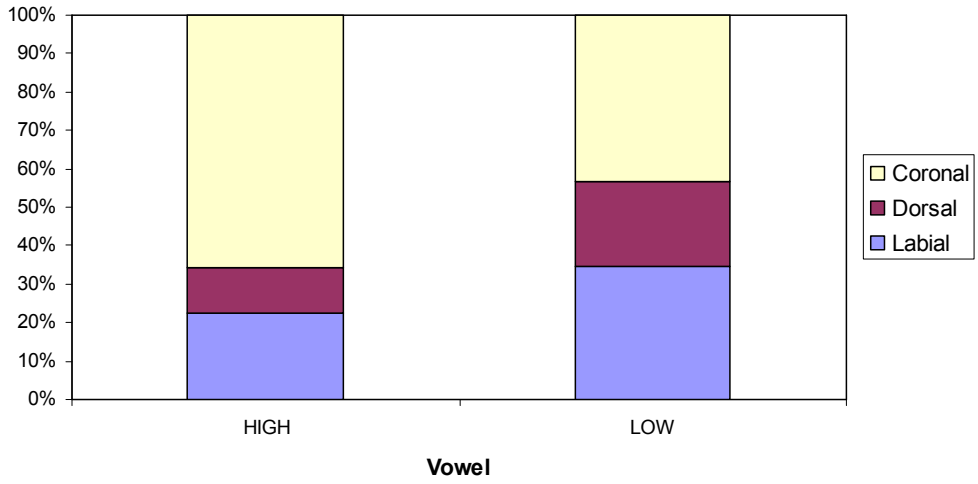
Cramer's V shows that in this case the effect size is larger in CELEX than in CMU. Figure 4 demonstrates the alignment between consonant place and vowel height in CELEX. We can see that high vowels combine more often with coronal consonants, while low vowels, on the contrary, combine more often with dorsal and labial consonants.

But for the coda-vowel height relationship the association between the two variables is again stronger in CMU than in CELEX:

(9) CMU: $\chi^2(2, N = 83{,}798) = 1914.354$, $p < .001$, Cramer's V = 0.151.
CELEX: $\chi^2(2, N = 25{,}888) = 56.395$, $p < .001$, Cramer's V = 0.047.

The Cramer's V values show that the association between coda place and vowel height is weaker than between onset place and vowel height.

Figure 4: Distribution of onset consonants across LOW and HIGH vowels in CELEX



Vowel-stress alignment

The results showed that the OCP in stressed syllables may be weaker than in unstressed syllables, at least for dorsal and labial co-occurrences. At the same time, the OCP appears to be weaker with low or mid vowels than with high or reduced vowels. If low and mid vowels were more common in stressed vowels than in unstressed it would work as a link between the two patterns. We examined the vowel distribution across the two syllable types. Tables 16 and 17 represent O/E values for the low and high vowels in stressed and unstressed syllables.

Table 16: The distribution of vowels in stressed and unstressed CVC syllables of CMU.

		Syllable type		
		Stressed	Unstressed	
Vowel quality	Low vowels	19404.00	3688.00	Observed
		8003.58	15088.42	Expected
		2.42	0.24	O/E
	High vowels and schwa	9640.00	51066.00	Observed
		21040.42	39665.58	Expected
		0.46	**1.29**	O/E

Table 17: The distribution of vowels in stressed and unstressed CVC syllables of CELEX.

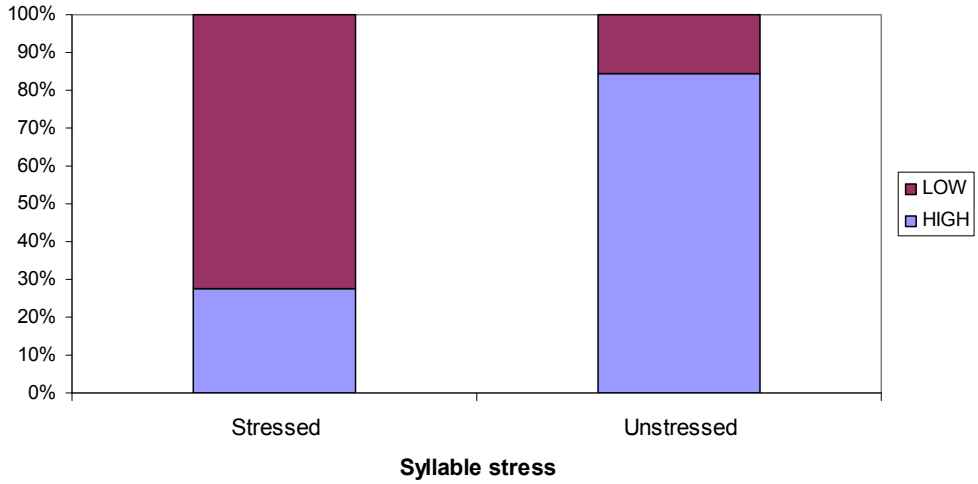
		Syllable type		
		Stressed	Unstressed	
Vowel quality	Low vowels	7480.00	2412.00	Observed
		3952.90	5939.10	Expected
		1.89	0.41	O/E
	High vowels and schwa	2865.00	13131.00	Observed
		6392.10	9603.90	Expected
		0.45	**1.37**	O/E

The results showed that low and mid vowels were over-represented in stressed syllables, while high and reduced vowels were over-represented in unstressed syllables. The Chi-square test for the association between vowel height and stress was highly significant, with the highest Phi and Cramer's V values.

(10) CMU: $\chi^2(1, N = 83{,}798) = 34306.56$, $p < .001$, Phi and Cramer's V = 0.640.
CELEX: $\chi^2(1, N = 25{,}888) = 8483.39$, $p < .001$, Phi and Cramer's V = 0.572.

Figure 5 illustrates this relationship. Stressed syllables contain more low vowels than high vowels, but in unstressed syllables high vowels prevail.

Figure 5: Distribution of HIGH and LOW vowels across stressed and unstressed syllables of CELEX.



The regression analysis

A multiple linear regression analysis was performed to evaluate the contribution of each factor to the well-formedness and frequency of the syllable type. The natural logarithm of the observed frequency was the dependent variable and a natural logarithm of the expected frequency was used as one of the independent factors. A number of categorical factors were again defined to estimate the strength of the OCP effect, the strength of the alignment between the factors, and the potential effect of vowel height on the strength of the OCP. For example, an OCP factor assigns '1' to each syllable type that violates OCP and '0' to each syllable type that does not violate OCP.

It also became apparent that high vowels in stressed syllables and low vowels in unstressed syllables are under-represented in the two datasets. To test these tendencies, two variables were constructed: Stressed/High and Unstressed/Low. Each factor assigns '1' to all syllables containing either combination and '0' to the syllables that do not. Analogous variables were constructed for the combinations of consonant place and vowel height or stress, which were expected to affect the frequency of the syllables, based on the previously presented results. For example, the combinations of coronal consonants with low vowels, and the combinations of labials or dorsal consonants with high vowels, are each expected to occur at lower frequencies. Similarly, coronal consonants in stressed syllables, and labial or dorsal consonants in unstressed syllables, are each less likely to occur. However, only one of these two sets of variables can be entered into the regression, because they are highly correlated through the vowel-stress connection. Therefore the regression is tested with each set of factors separately to determine which one produces the best fit.

The cases included in the analysis represent all possible combinations of the investigated categories. The factors we were interested in are onset place (coronal, dorsal, and labial), coda place (coronal, dorsal, and labial), stress (stressed and unstressed), and vowel height (high and mid/low). When combined these factors produce 36 possible syllable types, for example, LLHS (labial-labial, high vowel, stressed) or DCLU (dorsal-coronal, low vowel, unstressed).The best fit for CELEX data was obtained with a combination of the following factors:

Table 18: Linear regression for CELEX CVC syllables.

Factors	*Coefficients*	*SE*	*t*	*p-value*	*Tolerance*
Expected	0.856	0.088	9.689	< 0.001	0.965
OCP	-0.738	0.16	-4.599	< 0.001	0.989
Stressed/High	-0.830	0.2	-4.151	< 0.001	0.754
Unstressed/Low	-1.513	0.199	-7.591	< 0.001	0.759
Labial-Dorsal onset/High	-0.660	0.196	-3.372	< 0.01	0.664
Labial-Dorsal coda/High	-0.477	0.196	-2.43	< 0.05	0.660

The R value for this regression is 0.943 ($F(6, 35) = 38.689$, $p < 0.001$). The factors connecting *unmarked* place (coronals) and low vowels were not significant. Another factor that did not reach significance and is not presented in the table above is the factor testing whether OCP combined with high vowels reduced the frequency compared to OCP combined with low vowels (distance effect).

All the variables, as expected, had negative coefficients, as they assigned a value '1' to every combination that was expected to *lower* the observed frequency. The coefficients and their *p* values were higher for the OCP and stress-vowel factors than for the factors connecting place and vowel. Tolerance value was high for all the factors, which shows that collinearity was not a problem for this analysis.

The analysis showed that the frequency of the syllable was significantly reduced if it violated the OCP restriction and if it violated the vowel-stress alignment restriction. If the syllable combined a labial or dorsal consonant with a high or reduced vowel its frequency was also significantly reduced.

Table 19 presents the best fitting model for the CMU data.

Table 19: Linear regression for CMU CVC syllables.

Factors	*Coefficients*	*SE*	*t*	*p-value*	*Tolerance*
Expected	0.857	0.077	11.120	< 0.001	0.968
OCP	-0.961	0.204	-4.709	< 0.001	0.997
Stressed/High	-1.708	0.254	-6.723	< 0.001	0.762
Unstressed/Low	-1.169	0.257	-4.557	< 0.001	0.747
Labia-Dorsal onset /Unstressed	-0.820	0.251	-3.268	< 0.01	0.660
Labia-Dorsal coda/Unstressed	-0.737	0.250	-2.947	< 0.01	0.664

The R value for this regression is 0.945 ($F(6, 35) = 13.515$, $p < 0.001$). This model is remarkably similar to the one fitted to the CELEX data. But there are a few differences as well. First of all, although for CELEX data place-*vowel* factors were the best predictors, in the case of CMU place-*stress* factors produced better results. At the same time, in both datasets, only the factors connecting *marked* place (labials and dorsals) with unstressed syllables were significant. Those that connected unmarked coronal consonants to stressed syllables did not reach significance. Again for both datasets, the highest coefficients were obtained for the OCP and stress-vowel factors. Again, as expected, all coefficients were negative, which meant that they significantly *lowered* the observed frequency.

The results of the regression showed that the syllables in the CMU were significantly reduced in frequency if they violated the OCP requirement, violated the vowel-stress alignment requirement, or contained labial or dorsal consonants in unstressed syllables.

Discussion

The initial hypothesis that the OCP effect is active not only in the monosyllabic words of English but in all CVC syllables was confirmed. Combinations of homorganic consonants in onsets and codas of the CVC syllables were generally under-represented compared to heterorganic combinations. This difference was statistically significant for both datasets: the OCP violations significantly reduced frequency of the syllables.

The results suggest that the OCP is avoided more rigorously in the CVC syllables than in the CVC words. The pattern of OCP violations in CVC syllables was closer to that reported in Berkley's study than the pattern found here for monosyllabic (but not monomorphemic) words. Monosyllabic words are likely to be more affected by a small amount of unusual or rare words, foreign borrowings, and proper names. Such words are not necessarily representative of the phonotactic patterns in the lexicon. Inclusion of all CVC syllables in the analysis reduces the impact of such words and increases the contribution of the more frequent well-formed syllables of English. Therefore, the results obtained from all CVC syllables in the two dictionaries may be less confounded by the outliers and more representative of the actual statistical patterns in the lexicon.

In particular, in the CVC words of CMU and CELEX dorsal co-occurrences had higher O/E values than in Berkley's results and patterned with coronals rather than with labials. This was confirmed by the statistical analysis, which showed that the prohibition against the OCP-labial was significantly stronger than against either the OCP-dorsal or the OCP-coronal. In the CVC syllables, on the other hand, the O/E values for the dorsal co-occurrences moved closer to the labials. The statistics showed that in this dataset the OCP-dorsal often patterned with the OCP-labial and differed from the OCP-coronal (the prohibition against the OCP-labial and the OCP-dorsal was stronger than against the OCP-coronal). And yet the O/E values for dorsal and coronal co-occurrences in the CVC syllables of CMU and CELEX are still higher than in Berkley's study. This difference may be attributed to the fact that our data were not morphologically uniform. Crucially, most discrepancies with Berkley's results are found for dorsal or coronal consonants, the ones most actively involved in the formation of English bound morphemes. In fact, the most frequent morphemes of English, especially the inflectional morphemes, often include coronals. One of the most common inflections, /ɪŋ/, includes a dorsal consonant. Therefore, higher O/E values for co-occurrences of coronals and co-occurrences of dorsals in our results are possibly due to the presence of morphologically complex words in the datasets.

The distance effects in CVC syllables showed more consistency than in monosyllabic words. There is a tendency for the strength of the OCP to decrease with long vowels and diphthongs, compared to short vowels. The difference between the short vowels and diphthongs reached significance in the regression analysis for CELEX.

The effect of the vowel height on the magnitude of the OCP was not confirmed. While the O/E values suggested that the OCP-labial and the OCP-dorsal may be weaker with low vowels, the difference was in the opposite direction for the OCP-coronal. This partial effect of vowel height failed to reach significance in the regression analysis.

The effect of stress on the magnitude of the OCP was similar to that of the vowel height. According to the differences in the O/E values it appeared that the OCP-labial and the OCP-dorsal are stronger in unstressed syllables than in stressed syllables, but the opposite was found for the OCP-coronal. This difference in the magnitude of the OCP effect imposed by stress was not significant in the regression analysis.

Crucially, the reversal of the predicted direction for coronals reveals another tendency in the data that is much stronger than the hypothesized distance effect. Coronal co-occurrences consistently appear to be over-represented in unstressed syllables and with high vowels, while the distance effect predicts that they should be under-represented. Not only co-occurrences show this behavior. The distributional patterns in the data revealed a tendency for stressed syllables and low/mid vowels to combine with dorsal and labial consonants; unlike unstressed syllables and high/reduced vowels, which prefer to pair with coronal consonants. This result is evidence for prominence alignment between consonant place, vowel height, and stress. Based on the distance effect, dorsal-vowel-dorsal syllables should have higher O/E values with low vowels and in stressed syllables because OCP is more tolerated with low vowels. But prominence alignment also predicts that dorsals are more frequent onsets and codas in stressed syllables and with low vowels. In this case the two possible explanations go in the same direction and it is difficult to tease them apart. The case of coronals resolves the issue. While according to the distance effect coronal-vowel-coronal syllables are to be more tolerated with low vowels and in stressed syllables, prominence alignment requires coronals to combine with high vowels and unstressed syllables. The data patterns with the prominence alignment, and either the distance effect does not apply in this case or it is not powerful enough to overcome the requirements of the prominence alignment. In addition, the distance effect never reached significance in the regression analysis, while the prominence alignment factors proved to be significant predictors of the syllables' frequency.

The alignment between consonant place and stress and between consonant place and vowel height are so similar because they are connected through a third alignment in the phonotactics of English syllables: the alignment between vowel height and stress. Stressed syllables contain more low and mid vowels, and unstressed syllables contain more high and reduced vowels. This is the most powerful factor, which determines the well-formedness of the CVC syllable. Violation of vowel-stress alignment requirements significantly reduces the frequency of the syllable. The OCP violation is the second-strongest factor in the regression analysis. The factors concerning the alignment between consonant place and vowel height or stress are third-strongest. Chi-square tests established that both onsets and codas interacted with stress/vowel height: labial and dorsal onsets and codas were preferred in stressed syllables, but coronal onsets and codas were preferred in unstressed syllables. We also observed that this connection was stronger for onsets than for codas. The regression showed, however, that only the connection between labial and dorsal consonants and stress/vowel height reached significance. The way it can be interpreted is that marked consonants -- labials and dorsals -- are prohibited in unstressed syllables or/and in combination with high vowels. There is no strong restriction against unmarked coronal consonants in stressed syllables. This finding strongly supports the positional neutralization view on the distribution of consonants in English lexicon: while stressed syllables preserve the identity of the consonants intact, unstressed syllables work towards the neutralization of consonants to the unmarked coronal place of articulation. In contrast, the vowel-stress alignment in English phonotactics is symmetric: it requires both positional

neutralization (high and reduced vowels in unstressed syllables) and positional augmentation (low and mid vowels in stressed syllables).

To summarize the results of the statistical analysis, the regression showed that the observed frequency of CVC syllables of English was significantly reduced if (a) the syllable was stressed but contained a high vowel, (b) a syllable was unstressed but contained a low vowel, (c) a syllable incurred an OCP violation, (d) a syllable was unstressed but had a dorsal or labial onset OR a labial or dorsal onset was followed by a high vowel, (e) a syllable was unstressed but contained a dorsal or labial coda OR a high vowel was followed by a dorsal or labial coda.

4. OT analysis

Constraints

First we define a set of constraints to be included in the OT grammar. Descriptive and statistical analyses suggest the generalizations to be expressed as constraints. The prohibition against homorganic consonants in the onset and the coda of the CVC is one of the candidates:

(11) **OCP** Consonants of the same place of articulation are prohibited in the onset and the coda of the same syllable.

Another set of constraints that need to be included in the grammar are the constraints working against non-harmonic alignment of stressed syllables with high vowels and unstressed with low vowels.

(12) ***x/a** Low vowels are prohibited in unstressed syllables
***X/i** High vowels are prohibited in stressed syllables

A final group of constraints the grammar needs are the constraints against non-harmonic alignment between consonant place and vowel height or stress.

(13) ***x/p** Labial and dorsal consonants are prohibited in unstressed syllables
***i/p** Labial and dorsal consonants are prohibited with high vowels

These constraints, just as the vowel-stress relating constraints, arise from the combination of three prominence scales: stressed syllables are more prominent then unstressed, low vowels are more prominent than high and reduced vowels, and labial and dorsal consonants are more prominent than coronal consonant.

The following summarizes the constraints, which will be included in the grammar:

(14) **OCP** Consonants of the same place of articulation are prohibited in onset and coda of the same syllable
***x/a** Low vowels are prohibited in unstressed syllables
***X/i** High vowels are prohibited in stressed syllables
***x/p_** Labial and dorsal onsets are prohibited in unstressed syllables
***x/_p** Labial and dorsal codas are prohibited in unstressed syllables
***p_/i** Labial and dorsal onsets are prohibited when followed by high vowels
***i/_p/** Labial and dorsal codas are prohibited with preceded by high vowels

An additional faithfulness constraint, FAITH, is added to each grammar to penalize any deviation from the faithful mapping.

Partially ordered grammars and gradient phonotactics

How can these constraints be used to capture the *gradient* restrictions on certain kinds of syllables in English phonotactics? Standard Optimality Theory is not capable of capturing gradient phenomena. Each of these constraints is categorical and, if ranked high in standard OT, will *completely* rule out the syllable type it is working against. The situation that we observe in English, however, is that the syllable types that violate these constraints are not completely banned from the language, but merely under-represented in comparison with their counterparts. Various proposals for modifying standard OT included soft or gradient constraints – constraints which can be violated *as often* or to such *a degree* that is required for a candidate to surface in the proportions observed in the lexicon (Berkley 1994, Frisch *et al.* 2004). Another proposal, known as Harmonic Grammar, involves lexically indexed constraints or weighted constraints (Pater and Coetzee 2005, Coetzee and Pater 2008). All these proposals require significant additions or modifications of the standard OT. While they do relatively well in accounting for gradient phonotactics in particular target languages, they have to sacrifice the simplicity and parsimony of the grammar. Treatments of gradient phonotactics, which allow the introduction of numerical values observed in the corpus to the model (Coetzee and Pater 2008), are very powerful descriptively but lack in explanatory value. They can model with precision quantitative patterns observed in the data but they make no predictions concerning the typological validity of the constructed grammar. Since these models rely on actual numbers derived from the data, their predictions can be easily reversed if the numbers are reversed. Therefore, they put no limitations on quantitative relationships observable in the languages.

The approach we explore here offers both descriptive precision and typological predictions. It relies only on the devices already available in standard OT. Frequency patterns in the data are modeled through constraint interaction, without introducing actual numbers into the grammar. It derives quantitative facts of gradient well-formedness by removing the limitations of strict constraint ranking. The approach known as Partially Ordered Grammars (Anttila 1998) was originally developed to deal with variation. The core idea of the proposal is that not all constraints have to be ranked in respect to each other in the grammar. When a matrix of all candidates, constraints and violations are submitted to an OT analysis the result is a *factorial typology*, a set of possible grammars resulting from all possible permutations of the given constraints. This technique is routinely used in theoretical phonology to test the predictions of each individual grammar against the typological patterns attested in the world's languages. Anttila proposed that a single language need not adhere to a unique constraint ranking. If some of the constraints in a language can be freely re-ranked the result is a number of grammars with different winning candidates, which gives rise to variation. A given variant is predicted to occur at a higher frequency if more of the possible grammars allow that candidate to win out over competing candidates.

The same approach is applied here to gradient phonotactics. The relative well-formedness of the candidates is calculated similarly to the relative frequencies of the variants. The model predicts that the more marked candidates will be the least well-formed and consequently the least frequent forms in the lexicon. The architecture of the model allows it to translate markedness

into frequency by outputting more marked candidates in a lower proportion than less marked candidates.

In addition to its descriptive adequacy, the model presented here provides typological predictions. The quantitative relationships between the candidates in the grammar are derived from their constraint violations, not directly from their frequencies in the data. These relationships cannot be reversed and as a result the model makes concrete and testable predictions about frequency patterns that are expected to emerge in languages. For example, certain syllable types are predicted to be universally more well-formed and more frequent than others. The existence of less well-formed syllable type in the language implies that that language will also feature more well-formed types.

Testing the grammar

Applying the model to the data at hand will reveal how closely the markedness relationships constructed by the grammar follow the quantitative patterns observed in the data. The relative markedness of outputs can be represented as an implicational hierarchy, where the more marked outputs entail the less marked outputs. In other words, if a grammar allows a more marked candidate to win, it is implied that the less marked candidate will win as well. This insures that less marked outputs win more frequently than more marked ones. In a successful grammar all entailments should follow the order of frequencies in the data. That is, where form A is more frequent than form B in the data, A will entail B in the grammar.

The task of finding all entailments in the factorial typology becomes increasingly difficult as the number of candidates and constraints grows. Factorial typologies can be computed with the help of OTSoft (Hayes, Tesar, and Zuraw 2003), and a T-order Generator (Anttila & Andrus 2006) then extracts all implicational hierarchies from that factorial typology (or directly from the Excel file containing candidates). The T-order Generator can also represent the entailments in the grammar graphically and evaluate the success of the grammar with the help of two measures: precision and recall. Precision indicates how many of the predicted quantitative relationships were observed, while recall indicates how many of the observed quantitative relationships were predicted. Less then perfect precision means some of the observed quantitative patterns were reversed in the grammar. Less then perfect recall suggests that the grammar constructs fewer relationships than are observed in the data. This happens when two forms differ in frequency but the grammar does not discriminate between them. Both measures range between 0 and 1, where 1 indicates perfect recall or precision.

Candidates for the grammar were all possible syllable types (the same as in the regression analysis). The combination of four factors: onset place (three levels), coda place (three levels), stress (two levels) and vowel height (two levels) produced 36 possible combinations. An observed and expected frequency, as well as an O/E ratio, was computed for each syllable type in CMU and CELEX. Constraints used for each grammar were selected based on the results of the regression:

(15) <u>For CMU:</u>

OCP	consonants of the same place of articulation are prohibited in onset and coda of the same syllable
***x/a**	low vowels are prohibited in unstressed syllables
***X/i**	high vowels are prohibited in stressed syllables
***x/p_**	marked onsets are prohibited in unstressed syllables

***x/_p** marked codas are prohibited in unstressed syllables
FAITH output the candidate without modifications

(16) For CELEX:
OCP consonants of the same place of articulation are prohibited in onset and coda of the same syllable
***x/a** low vowels are prohibited in unstressed syllables
***X/i** high vowels are prohibited in stressed syllables
***p_/i** marked consonants are prohibited when followed by high vowels
***i/_p** marked consonants are prohibited with preceded by high vowels
FAITH output the candidate without modifications

Tables 20 and 21 summarize the frequency counts for the syllable types in each dictionary, listed according to O/E value.

Table 20: Matching pattern of constraint violations with O/E values in CMU

Onset	Coda	Vowel	Stress	Example	Observed frequency	Expected frequency	O/E value
labial	labial	high	unstressed	[bi:f]	37	823.790	0.045
dorsal	dorsal	low	unstressed	[hæk]	28	340.032	0.082
labial	labial	low	unstressed	[bob]	38	313.362	0.121
coronal	coronal	low	unstressed	[tol]	992	6646.423	0.149
labial	labial	high	stressed	[bi:p]	106	436.975	0.243
dorsal	coronal	low	unstressed	[kon]	474	1845.666	0.257
dorsal	dorsal	high	stressed	[kɪk]	126	474.166	0.266
coronal	dorsal	low	unstressed	[leg]	414	1224.488	0.338
coronal	coronal	high	stressed	[lut]	3154	9268.269	0.340
coronal	labial	low	unstressed	[top]	217	630.302	0.344
labial	coronal	low	unstressed	[bel]	1147	3304.348	0.347
coronal	dorsal	high	stressed	[dɪg]	632	1707.517	0.370
labial	dorsal	low	unstressed	[peg]	230	608.769	0.378
dorsal	dorsal	high	unstressed	[kɪŋ]	410	893.901	0.459
labial	dorsal	high	stressed	[bIg]	405	848.913	0.477
dorsal	coronal	high	stressed	[kɪt]	1280	2573.734	0.497
dorsal	labial	high	stressed	[gɪv]	145	244.076	0.594
coronal	labial	high	stressed	[zɪp]	576	878.940	0.655
labial	coronal	high	stressed	[bul]	3216	4607.830	0.698
coronal	labial	high	unstressed	[tɪv]	1740	2299.263	0.757
dorsal	labial	low	unstressed	[kom]	148	175.030	0.846
labial	dorsal	high	unstressed	[wɪg]	1395	1600.377	0.872
dorsal	labial	high	unstressed	[kəm]	442	460.133	0.961
dorsal	coronal	high	unstressed	[gud]	5402	4852.025	1.113
labial	coronal	high	unstressed	[mɪs]	10401	8686.721	1.197
coronal	coronal	low	stressed	[lot]	5114	3525.564	1.451
coronal	coronal	high	unstressed	[tən]	25617	17472.621	1.466
coronal	dorsal	high	unstressed	[dɪŋ]	5622	3219.027	1.746
labial	labial	low	stressed	[mæm]	363	166.221	2.184
dorsal	dorsal	low	stressed	[ho:k]	401	180.368	2.223
coronal	dorsal	low	stressed	[dog]	1587	649.524	2.443

labial	dorsal	low	stressed	[bæk]	820	322.918	2.539
labial	coronal	low	stressed	[wel]	5315	1752.776	3.032
dorsal	coronal	low	stressed	[hot]	3403	979.025	3.476
coronal	labial	low	stressed	[dæm]	1549	334.341	4.633
dorsal	labial	low	stressed	[hæm]	852	92.844	9.177

Table 21: Matching pattern of constraint violation with O/E values in CELEX

Onset	Coda	Vowel	Stress	Example	Observed frequency	Expected frequency	O/E value
labial	labial	high	unstressed	[bi:f]	27	404.542	0.067
dorsal	dorsal	high	stressed	[kɪk]	40	209.855	0.191
coronal	coronal	low	unstressed	[tol]	453	2159.483	0.210
dorsal	dorsal	low	unstressed	[hæk]	50	194.982	0.256
coronal	coronal	high	stressed	[lut]	747	2324.196	0.321
coronal	labial	low	unstressed	[top]	177	529.234	0.334
coronal	labial	high	stressed	[kɪt]	217	637.393	0.340
coronal	dorsal	low	unstressed	[leg]	250	710.985	0.352
labial	labial	high	stressed	[bi:p]	95	269.252	0.353
dorsal	labial	high	stressed	[gɪv]	56	156.209	0.358
labial	labial	low	unstressed	[bob]	94	250.171	0.376
coronal	dorsal	high	stressed	[dɪg]	322	765.215	0.421
labial	dorsal	high	stressed	[bɪg]	202	361.720	0.558
dorsal	labial	low	unstressed	[kom]	89	145.138	0.613
labial	coronal	low	unstressed	[bel]	639	1020.796	0.626
dorsal	coronal	low	unstressed	[kon]	393	592.222	0.664
coronal	labial	high	stressed	[zɪp]	384	569.600	0.674
dorsal	labial	high	unstressed	[kəm]	170	234.698	0.724
labial	coronal	high	stressed	[bul]	802	1098.657	0.730
dorsal	dorsal	high	unstressed	[kɪŋ]	236	315.299	0.748
labial	dorsal	low	unstressed	[peg]	267	336.086	0.794
labial	dorsal	high	unstressed	[wɪg]	520	543.472	0.957
labial	coronal	high	unstressed	[mɪs]	1946	1650.693	1.179
dorsal	coronal	high	unstressed	[gud]	1155	957.661	1.206
coronal	coronal	low	stressed	[lot]	1780	1437.294	1.238
dorsal	dorsal	low	stressed	[ho:k]	183	129.775	1.410
labial	labial	low	stressed	[mæm]	241	166.507	1.447
coronal	coronal	high	unstressed	[tən]	5549	3492.023	1.589
coronal	labial	high	unstressed	[tɪv]	1574	949.649	1.657
coronal	dorsal	high	unstressed	[dɪŋ]	1954	1149.709	1.700
coronal	dorsal	low	stressed	[dog]	858	473.213	1.813
coronal	labial	low	stressed	[dæm]	771	352.244	2.189
labial	dorsal	low	stressed	[bæk]	532	223.689	2.378
labial	coronal	low	stressed	[wel]	1640	679.414	2.414
dorsal	coronal	low	stressed	[hot]	1123	394.167	2.849
dorsal	labial	low	stressed	[hæm]	352	96.600	3.644

Shaded areas of onset and codas columns indicate syllable types that violate OCP. Shaded areas of the vowel and stress columns indicate syllable types that violate vowel-stress constraints. These cases are almost completely limited to the upper part of the table, suggesting

that OCP and vowel-stress constraints play an important role in selecting the most under-represented syllable types.

When these candidates, their O/E values, unranked constraints, and violation patterns were submitted to a T-order generator, the following results were obtained:

(17)

DATA	RANKINGS	PRECISION	RECALL
CMU	none	0.85	0.54
CELEX	none	0.86	0.55

These numbers show a fairly good precision but a rather low recall. The high precision numbers indicate that there are relatively few cases of reversed relationships in our grammar, cases when one syllable type is more frequent then another in the data, but the grammar predicts the opposite. Low recall, on the other hand, means that there are many cases when there is a difference in O/E values between two candidates but, since both incur the same amount of constraint violations, the grammar does not distinguish them. This problem can be overcome in two ways. If the candidates that need to be distinguished violate the *same* constraints, then additional constraints have to be introduced. If the candidates have the same number of violations but from *different* constraints, these constraints can be ranked with respect to each other in order to distinguish between the candidates. Here, the second option is more appropriate.

Statistical analysis showed that not all factors made equals contributions to the model. Vowel-stress factors had the highest coefficient, with the prohibition against low vowels in unstressed syllables being stronger than the prohibition against high vowels in stressed syllables. The OCP factor followed the vowel-stress factor in terms of strength. Finally, the consonant-vowel factor and consonant-stress factor had the lowest coefficients in the model. At the same time, onset-related factors had higher coefficients than coda-related factors. This order of factor coefficients in the regression analysis motivated the following rankings among the markedness constraints in the grammar.

(18) For CMU:
*x/a >> *X/i >> OCP >> *x/p_ >> *x/_p

(19) For CELEX:
*x/a >> *X/i >> OCP >> *p_/i >> *i/_p

The grammars with these rankings produced the following results:

(20)

DATA	RANKINGS	PRECISION	RECALL
CMU	all	0.75	0.88
CELEX	all	0.74	0.87

Adding ranking led to a decrease in precision values but a significant increase in recall values, improving the overall fit of the model. Introducing a-priori rankings made the grammar more discriminating, while the number of reversals between the predicted and observed entailments increased: not all of the relationships predicted by the rankings are reflected in the data.

Figure 6: T-order for CELEX CVC syllables.

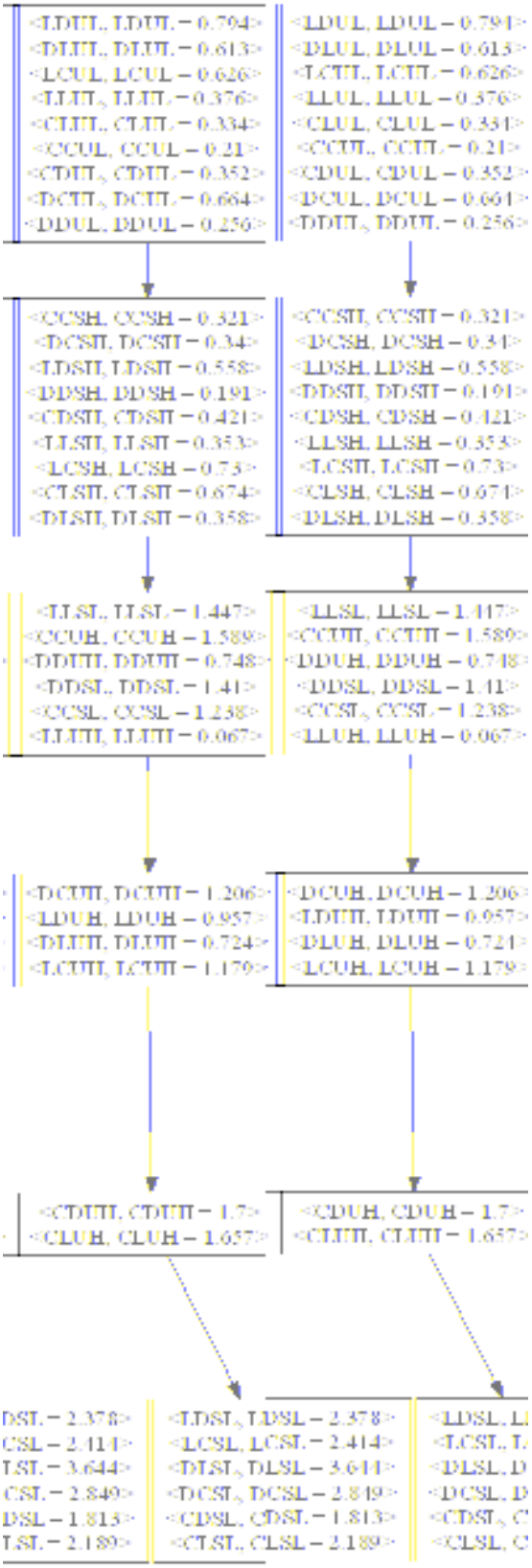


Figure 6, produced by the T-order Generator, illustrates the entailments established for the CELEX data. The grammar represented here is the best fitting grammar for the CELEX data, including all constraints and all rankings.

The input-output mappings at the top of the graph entail all the mappings below them (every arrow signifies an entailment). That is, the higher it is on the graph, the more marked the output is, meaning that it violates a higher ranked markedness constraint. The outputs at the top of the graph are the ones predicted to be the least well-formed and the most under-represented syllable types in the lexicon. Since each output is supplemented by its O/E value we can trace the increase in O/E values from the top of the graph to the bottom.

However, the increase in O/E values from the top to bottom of the graph is not in perfect order: there are cases of reversed relationships and outputs with very different O/E values being lumped together in the same box. These are indications that the grammar does not model the relationships in the data perfectly. It is possible, though, that perfect recall and precision may be neither achievable nor necessary. Perfect recall would mean that the grammar can discriminate between *every single* candidate, since each candidate has its own unique O/E value. However, a theory that argues for a difference in markedness between every single candidate is not particularly useful and would require a great number of otherwise unnecessary constraints.

On the other hand, perfect precision would mean that the quantitative relationships specified by the grammar are never reversed in the actual data. Any attempt to achieve perfect precision implies a belief that every observed quantitative relationship is *meaningful* and that there are no *accidental* reversals. This is again impractical, knowing that linguistic data is noisy and inevitably includes some accidental reversals and superfluous relationships. Therefore, I argue that the proposed model provides a reasonably close fit to the data.

5. General discussion and conclusion

This study makes a number of contributions to the understanding of the gradient phonotactics of English. First, it shows that the gradient restriction against the co-occurrences of the homorganic consonants in English operates in all CVC syllables. Previously this was established only for monomorphemic monosyllabic words (Berkley 1994).

Second, extending the analysis from monosyllabic words to syllables led to the discovery of a number of additional factors which play an important role in gradient phonotactics of English. Alongside the OCP restriction, a requirement for the prominence alignment between consonant place, vowel height, and stress define the well-formedness of the CVC syllables in English. It was established that stressed syllables tend to contain more marked (labial and dorsal) consonants and more prominent low vowels. Unstressed syllables tend to carry more unmarked (coronal) consonants and less prominent high or reduced vowels.

The strength of the OCP and harmonic alignment in the relative well-formedness of syllable types was examined in a series of Chi-square tests and in the multiple regression analyses. It was found that the restriction against prominent vowels in non-prominent positions (low vowels in unstressed syllable) and against non-prominent vowels in prominent positions (high vowels in stressed syllables) is the strongest factor in the phonotactics of English CVC syllables, followed by the restriction against the OCP violations. The requirement for the prominence alignment between consonant place and stress or vowel height was the weakest among the significant factors. This effect was stronger for onsets then for codas. A prohibition against unmarked consonants in prominent positions (in stressed syllables or with low vowels) did not reach significance in the regression analysis.

It follows from these results that an ideal stressed CVC syllable of English contains a low or a mid vowel and avoids OCP violations. An ideal unstressed syllable contains a high or reduced vowel and a coronal coda and onset. An OCP violation is less crucial in the latter case than the prominence alignment requirement. The rest of the syllables receive intermediate well-formedness ratings according to the number of restrictions they violate.

This result suggests that in unstressed syllables vowels tend to neutralize to high and reduced members of the inventory, while in stressed syllables they augment to more prominent low and mid vowels. Consonants on the other hand, neutralize to unmarked coronals in unstressed syllables, but the preference towards more marked labials and dorsals in stressed syllables is not so pronounced.

Augmentation in prominent positions can be viewed as another kind of neutralization, because as a result it reduces the inventory of the segments that can appear in this position. For example, the inventory of vowels in prominent stressed syllables in English is largely limited to low and mid vowels. Consonants in prominent positions, on the other hand, do not show this behavior. By failing to augment to labials and dorsals in stressed syllables they work towards maintaining a larger diversity of the syllable types in the prominent, stressed, position.

This divergent behavior of vowels and consonants in stressed syllables may help stressed syllables fulfill their specific functions in the communication process. Their acoustic salience makes them perfect candidates for an increased informational load. Their function as the main information units calls for an increased acoustic salience. A restriction to low and mid vowels in stressed syllables provides the necessary acoustic prominence and the extended consonantal inventory accommodates the increased informational load.

Neither neutralization in non-prominent positions nor augmentation in prominent positions is unusual. The case of vowel sonority and stress alignment is well attested and researched (cf. Kenstowicz 1994, de Lacy 1997, 1999, 2001). Stress is often attracted to low and mid vowels, and vowels in stressed syllables can undergo modifications to meet prominence requirements. Prominence related processes, however, have been traditionally limited to the segmental level and theorized not to affect subsegmental features. De Lacy (2001) explicitly states that the constraints of the kind exploited in this study (e.g., *X/t, no coronals in stressed syllables) should not exist due to the fact that no languages are attested where certain place features are prohibited in prominent positions. However, connections such as this do emerge as weak but significant effects in the present study. There is clearly a statistical tendency in English CVC syllables to avoid labial and dorsal consonants in unstressed syllables. It is possible that this tendency is not strong enough to become fully phonologized in alternations. More independent evidence confirming this connection is required to fully establish the prominence alignment between consonant place and prominence features of the syllables.

Another question is whether there is any phonetic basis for the consonantal prominence scale. Are labial and dorsal consonants phonetically more prominent than coronal consonants? Articulatorily, labial and dorsal gestures involve bigger articulators and consequently require relatively more time and effort than coronal gestures. Increased time and effort may result in a more salient acoustic output – a longer and more intense sound.

Following Anttila (2008), the observed statistical patterns in the lexicon were modeled in Optimality Theory with the help of a partially ordered grammar and T-orders. This approach was chosen for its theoretical parsimony and predictive power. It not only models the quantitative patterns in the data but also provides concrete and testable typological predictions. The study

also shows the advantages of supplementing a descriptive approach with a statistical analysis. The results of the regression were used in the construction of the OT constraints and imposing the appropriate rankings among them. The factors that emerged as significant predictors of the syllables' frequency in the regression became the constraints in the OT grammar, and the factors with the highest coefficients in the regression were ranked higher in the OT grammar, resulting in the significant improvement of the model fit. In addition, the regression analysis revealed an important generalization: vowels participate in both positional neutralization and augmentation but consonants only participate in neutralization.

The data suggest a number of additional phenomena which were not addressed in this study and which are ideal for the future work. The most important remaining question is the possible effect of the inflectional and derivational morphology on the statistical patterns in the data. One indication of this possible effect is that the OCP restrictions in the data examined here do not appear as strict as those reported by Berkley (1994). A possible explanation is that the OCP effect is weaker or perhaps non-existent across morphological boundaries. Inclusion of morphologically complex items in the dataset could raise the overall tolerance of the OCP violations. Another indication is that many of the established patterns were more robust in CMU than in CELEX dictionary, especially the connection between consonant place and stress. This suggests that inflectional endings, present in CMU but not CELEX, could be responsible for boosting certain quantitative patterns. It is necessary to separate affixes from the roots to verify this prediction. The expectation is that observed patterns will still hold in the roots alone, meaning that inflectional morphology only magnifies the effect rather than creates it.

In conclusion, this paper has expanded our understanding of the role of the OCP in English phonotactics and provided evidence that prominence alignment between vowel height, stress, and consonant place is an equally important factor in the phonotactics of English syllables.

References

Anttila, Arto. 1998. Deriving variation from grammar. In Franse Hinskens, Roeland Van Hout and W. Leo Wetzels (eds.), *Variation, Change, and Phonological Theory*, John Benjamins, Amsterdam, pp. 35-68.

Anttila, Arto. 2008. Gradient phonotactics and the Complexity Hypothesis. To appear in *Natural Language and Linguistic Theory.*

Anttila, Arto and Curtis Andrus. 2006. T-Orders. Stanford University, ms. Available online at http://www.stanford.edu/~anttila/research/software.html.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX Lexical Database (Release 2). Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

Berkley, Deborah. 1994b. The OCP and gradient data. *Studies in the Linguistic Sciences* 24.59–72.

Blumenfeld, Lev. 2006. Constraints on phonological interactions. Stanford University: PhD Dissertation. Available on ROA.

Coetzee, Andries. 2004. What it means to be a loser: non-optimal candidates in Optimality Theory. University of Massachusetts, Amherst: Ph.D. Dissertation.

Coetzee, Andries, & Pater, Joe. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. To appear in *Natural Language and Linguistic Theory.*

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe. 2004. Similarity avoidance and the OCP. Natural Language and Linguistic Theory 22(1).179-228.

Goldsmith, John. 1976. Autosegmental Phonology. Cambridge, MA: MIT dissertation.

Greenberg, Joseph. 1950. The patterning of root morphemes in Semitic. Word 6. 162-81.

Hayes, Bruce, Bruce Tesar, and Kie Zuraw. 2003. OTSoft 2.1, software package. Available online at http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Kenstowicz, Michael. 1994. Sonority-driven stress. MIT: Ms. ROA #33.

de Lacy, Paul. 1997. Prosodic categorization, MA Thesis, University of Auckland.

de Lacy, Paul. 1999. Tone and prominence. University of Massachusetts, Amherst, ms.

de Lacy, Paul. 2001. Markedness in prominent positions. In Ora Matushansky, Albert Costa, Javier Martin-Gonzalez, Lance Nathan, and Adam Szczegielniak (eds.) HUMIT 2000, MIT Working Papers in Linguistics 40. Cambridge, MA: MITWPL, pp.53-66. ROA #432.

Leben, William. 1973. Suprasegmental phonology. Cambridge, MA: MIT dissertation.

McCarthy, John J. 1986a. OCP effects: Gemination and antigemination. Linguistic Inquiry 17.207–263.

McCarth, John. 1988. Feature geometry and dependency: a review. *Phonetica* 43. 84-108.

Odden, David. 1986. On the role of the Obligatory Contour Principle in phonological theory. Language 62.353–383.

Pater, Joe, & Coetzee, Andries. 2005. Lexically Specific Constraints: Gradience, Learnability, and Perception. In *Proceedings of the 3rd Seoul International Conference on Phonology*. Seoul: The Phonology-Morphology Circle of Korea. 85-119.

Prince, Alan & Smolensky, Paul. 2004. Optimality Theory: Constraint interaction in generative grammar. Malden, MA: Blackwell. [Revision of 1993 technical report, RUCCS. Available on Rutgers Optimality Archive, ROA-537].

Schein, Barry & Steriade, Donca. 1986. On Geminates. *Linguistic Inquiry* 17. 691-744.

Weide, Robert. 1998. The CMU pronunciation dictionary (Release 0.6). Carnegie Mellon University. Available online at http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Yip, Moira. 1988. The Obligatory Contour Principle and phonological rules: a loss of identity. Linguistic Inquiry 19. 65–100.