Computational Linguistics in the Netherlands Journal 3 (2013) 174-192 Submitted 06/2013; Published 12/2013

A linear model for exploring types of vowel harmony

Lili Szabó LILIS@COLI.UNI-SAARLAND.DE
Çağrı Çöltekin C.COLTEKIN@RUG.NL

Saarland University
University of Groningen

Abstract

In this paper, we present a computational/corpus study of vowel harmony, which is a phonotactic constraint that influences the choice of vowels *within* a word. We argue that languages with vowel harmony can be described better by statistical models predicting co-occurrence of vowels from their articulatory-phonetic features in comparison to languages that do not exhibit vowel harmony. We use a simple linear model that predicts co-occurrence of the vowels based on their articulatory features. Using child-directed speech and larger corpora of written text in four languages (Hungarian, Turkish, Dutch and English), we show that model fit is better for languages with vowel harmony compared to languages without vowel harmony. Furthermore our model also allows investigation of complex types of vowel harmony based on the phonetic features and their interactions. The aim of this study is to provide an exploratory tool for detecting and characterizing the vowel harmony process quantitatively in a language.

1. Introduction

Vowel harmony (VH) is a phonological process observed in some languages where vowels in a word harmonize according to some articulatory or phonetic feature. It is a non-local dependency between two vowels with some intervening segments, typically consonants but sometimes vowels, in between. In VH-languages vowels within a word agree in one or more of their articulatory features, e.g. backness, height and roundedness. Vowel harmony is observed in the roots of words as well as being part of the morphophonological process. That is, in languages with vowel harmony we find most of the word roots containing vowels that harmonize. Furthermore, the choice of allomorphs of the affixes that are attached to the root is affected by vowel harmony.

Example (1) below shows how VH influences suffixation in Hungarian and Turkish. In both languages the vowel in the suffix is determined by the [±front] (or [±back]) feature of the final vowel of the stem, and the appropriate allomorph in the dative morpheme gets selected.

	Hungarian		Turkish	
(1)	almá-n**a**k	'apple-DAT'	elma-y**a**	'apple-DAT'
	remek-n**e**k	'wonderful-DAT'	güzel-**e**	'beautiful-DAT'

Palatal (front–back) harmony is not the only type of harmony observed in world's languages. Other features which can be subject to vowel harmony include *roundedness*, *height*, *tongue position* and *nasalization*. Vowel harmony does not always present itself as a simple agreement of vowels over a single dimension. In some languages vowel harmony spans over more than one feature, or dimensions, and these features may interact. Furthermore, in some languages, some vowels may be transparent. A transparent vowel is not affected by or does not affect the choice of neighboring vowels.

Another aspect of vowel harmony that makes it interesting for computational and quantitative studies, is that VH is a strong tendency rather than a strict constraint. VH is violable and often violated. For example, the Turkish word *elma* 'apple' in (1) above, does not obey vowel harmony, as the first vowel in the word is a front vowel and the second vowel is a back vowel. As our experiments reported below will also confirm, vowel harmony is not a strict constraint, but a strong tendency. As such, it is best characterized using statistical methods. Studying VH in naturally occurring language data using computational methods may allow us to discover tendencies and sub-regularities that were not noted earlier.

Vowel harmony has been an interest to linguists, as it is part of the grammar of the languages that exhibit it. Correct characterization of the grammars of these languages is not possible without taking the vowel harmony into account. Besides the theoretical interest, understanding vowel harmony is also important for understanding child language acquisition. Vowel harmony is by no means a universal feature of natural languages. Children have to learn whether the input language exhibits vowel harmony, and, if it does, the type of vowel harmony present the input language. Once children learned the particular vowel harmony constraints that exist in their language, they can also put it in use for learning other aspects of the language, such as speech segmentation (Suomi et al. 1997, van Kampen et al. 2008, Ketrez 2013).

Studies of vowel harmony using computational techniques are relatively scarce, and the studies to date focus on unsupervised learners, typically investigating vowel harmony within a single feature dimension. As vowel harmony with multiple dimensions of classification of vowels, and the interaction between these dimensions is even less studied in the literature, our goal here is to fill this gap, by looking for the interaction between different types of VH. We do this by presenting experiments on naturally occurring linguistic data. We use conventional general linear models to investigate the relationship between multiple dimensions of articulatory-acoustic features of vowels, and their co-occurrence within a word. It has to be emphasized that these statistical models are not simulating vowel harmony, but detecting and describing them.

We test our models on four languages: Dutch, English, Hungarian and Turkish. Our VH-languages are Hungarian and Turkish; both exhibiting more than one type of VH. For control languages (non-VH languages) we use Dutch and English, in order to see that the effects of VH are not accidental, but specific only to VH-languages. The main results we present are from child-directed parts of CHILDES database (MacWhinney and Snow 1985). We also use larger, written language corpora to confirm our findings (Quasthoff et al. 2006).

Even though we intend to extend this work to model child language acquisition in the future, we stress that the work presented here is not intended as a model of how children learn or use vowel harmony during language acquisition. The contribution of the current work to child language acquisition research is through investigation of the input, by showing availability of cues in the input that may aid language acquisition.

The rest of the paper is organized as follows. In Section 2 we summarize the relevant work done in modeling vowel harmony. Section 3 describes the vowel inventory of the languages at investigation. Section 4 describes the method and the experiments, and reports the results. We discuss our findings and conclude in Section 6.

2. Related work

Early computational work addressing vowel harmony includes a connectionist system of learning vowel harmony of Hungarian by Hare (1990), and models of Turkish vowel harmony using Optimality Theory (Kirchner 1993), Boltzmann machines (Bellgrad 1993), and by constraint learning (Ellison 1994). Besides these modeling studies, Ketrez (2013) examines VH from the aspect of language acquisition, and shows that VH can be helpful in word segmentation for a language learner of a VH-language. Even though she does not present a model of vowel harmony, the study is relevant to our discussion here because of use of simple statistics and the multi-lingual nature of the study. In this section, we will discuss two recent modeling works, Goldsmith and Riggle (2012) and Baker (2009), in more detail.

Recent computational and quantitative research on vowel harmony mainly employs unsupervised machine learning approaches. Goldsmith and Riggle (2012) present a model of VH for Finnish. Their

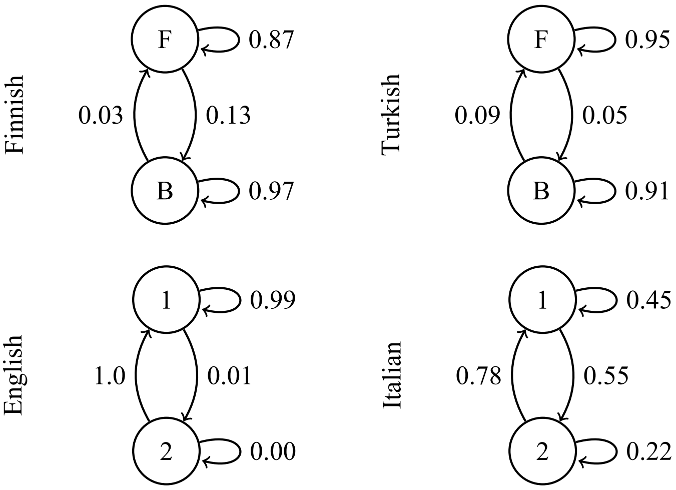


Figure 1: Two-state HMMs reported in Baker (2009). For VH-languages Finnish and Turkish the probability of transitions (marked on the edges) between the states are low, while probability of staying in the same state is high. The HMM learned from English data seems to prefer a single state, while the HMM for Italian has higher probabilities for state transitions. Note that the state labels front (F) and back (B) are assigned manually. Since this is the output of a unsupervised algorithm, the correspondence between the states and the vowel classes cannot be established automatically.

baseline models are simple unigram and bigram models on segments.[1] Their target model, called the Boltzmann model, is similar to the maximum entropy-like models often used in computational linguistic research. In their Boltzmann model, they make use of unigrams, bigrams and non-adjacent vowel bigrams, hence also allowing the model to learn about frequent non-adjacent vowel pairs. They first investigate whether the Boltzmann model yields an information theoretical gain—or equivalently assigns a higher probability to the data—compared to a model that makes use of only phoneme-unigram or -bigrams. Surprisingly, the Boltzmann model does not show an improvement over the simple phoneme-bigram model for Finnish. However, they show that the final cost assigned under minimum description length by the Boltzmann model to the complete corpus is lower than the cost assigned by the earlier bigram model.

In another recent study, Baker (2009) presents two different models on multiple languages. As well as the Boltzmann model of Goldsmith and Riggle (2012), Baker (2009) also reports experiments with Hidden Markov Models (HMMs) trained using the Baum-Welch algorithm (Manning and Schütze 1999, ch. 9). He reports experiments with both HMMs and Boltzmann models on multiple languages: Finnish, Turkish, English and Italian. Finnish and Turkish exhibit VH, while English and Italian do not. Similar to Goldsmith and Riggle (2012) he finds that the Boltzmann model with a vowel tier fits the data better for the languages that exhibit vowel harmony, while no improvement was observed for the languages without vowel harmony.

The experiments with HMMs are more interesting for our purposes here since inspection of the HMM learned from the data may allow one to gain further insights in the vowel harmony process. Example HMMs for VH and non-VH languages from this study are presented in Figure 1. The interpretation of such an HMM is relatively straightforward. For both VH-languages the probability of transition between the states is lower than probability of staying in the same state. Figure 1

[1] In this paper, we use the terms ‘phoneme’ and segment interchangeably, since the work discussed in this section, as well as our own experiments make use of idealized transcriptions without any phonetic variation.

shows two alternatives for non-VH languages. The model for English seems to prefer a single state. In other words, the model cannot find two discernible vowel classes for English. On the other hand, the HMM for Italian seems to admit two vowel classes,[2] but the probability of state switches is higher than the probability of staying in the same state. Baker (2009) also experiments with HMMs with higher numbers of states. He demonstrates that a four-state HMM can learn both front–back and roundedness harmony in Turkish.

HMMs are useful tools for investigating vowel harmony, since they can show the learnability of the phenomenon from the input only, and inspecting the resulting HMMs can provide further insights about the type of vowel harmony in the input language. Although HMMs can be trained using unsupervised algorithms, the number of states for HMMs should be chosen by the experimenter, and interpretation of the results may be difficult for HMMs with large number of states.

3. Languages

3.1 Hungarian

Hungarian exhibits palatal (front–back) vowel harmony (Kiss et al. 2003). In addition to that, roundedness also plays a role in some suffixation processes, but only for front vowels (as there are no unrounded back vowels in Hungarian). Lastly, a particular class of stems (so called opening stems) trigger the vowel of the suffixes or linking vowels to agree in height too.

The Hungarian standard vowel inventory consists of 14 vowel sounds, seven short and seven long vowel pairs. Except for the /a/–/ɒ:/ and /ɛ/–/eː/ pairs, there is only a small phonetic distinction between the short and long vowels: short vowels are slightly lower than their long counterparts (/i/–/iː/, /y/–/yː/, /u/–/uː/, /ø/–/øː/, /o/–/oː/). With a few exceptions[3] they are allophonic variations of the same phoneme. In our experimental setup, we collapse the short-long pairs with minimal spectral difference, and keep the ones with large spectral difference (/a/–/ɒ:/, /ɛ/–/eː/) as separate.

Three vowels, /ɛ/ and /i/–/iː/ are the so called *neutral* vowels which are transparent for palatal (backness) harmony. Although being front vowels according to their articulatory-acoustic features, they occur in front and back stems with equal frequency (Kiss et al. 2003), and don't influence the selection of correct allomorphs in suffixation processes, unlike in mixed words (stems with both front and back vowels—mostly loan words like *sofőr* [ʃoføːr] meaning 'chauffeur'— where always the last vowel in the word counts for selecting the allomorph for suffixation. Table 1 shows how the vowels are classified.

3.2 Turkish

Turkish has eight vowels which fill all slots of a vowel system classified according to features ±front, ±rounded, ±high. Table 2 presents all eight vowels and their articulatory-phonetic features.

Notable exceptions to Table 2 include three different allophones of /e/ with respect to height; long vowels /aː/, /uː/, /iː/ and /eː/ which share the phonetic features of their short counterparts and only occur in loan words; and palatalized (fronted) versions of /a/, /o/ and /u/ that only occur in loan words in the neighborhood of a palatalized consonant. Only the palatalized vowels are relevant to the vowel harmony process since when palatalized, these vowels are treated like their front counterparts.

Like Hungarian, Turkish also exhibits a palatal (front–back) vowel harmony. Additionally, Turkish vowels also harmonize according to their roundedness. There are no transparent vowels in Turkish. The palatal (frontness) harmony requires all vowels in a word to be either front or back

[2] What these classes correspond to is not evident without further inspection of the model and the data.

[3] Such as in [iro:] and [i:ro:].

	Front		Mid		Back	
	Rounded	Unrounded	Rounded	Unrounded	Rounded	Unrounded
High	/y/, /yː/	/i/, /iː/			/u/, /uː/	
Mid	/ø/, /øː/	/eː/,			/o/, /oː/	
Low-mid		/ɛ/				/ɒ/
Low				/aː/		

Table 1: Hungarian vowel chart and the phonetic classification of vowels. Although not strictly back, /aː/ always behaves like back vowel. The roundedness of /ɒ/ is also argued in the literature, phonetically it is rounded, but most of the times it behaves as unrounded. We treat it as unrounded here. Also, when we classify the height of the vowels mid and low-mid are merged into mid.

	Front		Back	
	Rounded	Unrounded	Rounded	Unrounded
High	/y/ 'ü'	/i/ 'i'	/u/ 'u'	/ɯ/ 'ı'
Non-high	/œ/ 'ö'	/e/ 'e'	/o/ 'o'	/a/ 'a'

Table 2: Turkish vowel chart and the phonetic classification of vowels. Each cell in the table presents the IPA symbol of most common allophone and the orthographic form of the vowel.

vowels. Roundedness harmony, on the other hand, requires round vowels to follow only round vowels, and unrounded vowels to follow rounded or unrounded vowels.

Native word roots follow both instances of the vowel harmony process with some exceptions. The major source of exceptions within the roots are the loan words. The morphophonological process is also quite regular with respect to vowel harmony, including the affixation of the loan words. The vowels in a suffix are consistently assimilated depending on the preceding vowel. The exceptions include (lexicalized) compounds and a few invariable suffixes.[4] Besides these exceptions that result in non-harmonic words, the choice of forms of some particles—most notably, the question particle *-mi/mı/mu/mü*— follow the VH process based on the last vowel of the preceding word.

3.3 Control languages - Dutch and English

In order to have a comparison with non-VH languages as well, we use Dutch and English as control languages. Both languages have simpler inflectional morphology compared to the VH languages in our study. Vowel reduction in unstressed syllables is a very common process in both Dutch and English.

The standard Dutch vowel inventory consists of 10 monophthongs and 6 diphthongs. 11 out of these classify as front vowels. All of the diphthongs are closing (the second sound is higher than the first one). The diphthongs do not move in front–back dimension.

The agreement of diphthongs (for both Dutch and English) is treated as follows in this study: when a diphthong is the first element of the bigram, the second (right side) sound of the diphthong is checked for agreement. On the other hand, if the diphthong is the second element of the bigram, its left part needs to agree with the first element of the bigram.

[4] For example, the imperfective suffix *-yor*, and the nominal relativizer *-ki*.

		sentences	words		vowels		
			tokens	types	tokens	lex. tokens	types
CHILDES	English	27482	111370	2555	111370	3097	15
	Dutch	23748	99023	4074	99023	4764	17
	Hungarian	24104	94162	9187	138715	9186	10
	Turkish	10206	33273	5388	58123	5387	14
Leipzig	English	96227	1920154	92889	2665600	170258	15
	Dutch	91952	1397413	99323	1995755	189250	19
	Hungarian	99684	1850713	216847	3102816	427357	10
	Turkish	97546	1306569	207789	2405691	409767	14

Table 3: The number of sentences, words and vowels in the corpora used in this study. For words and vowels we present both type and token counts. The column 'lex. tokens' for the vowels lists the token count of the vowels in the word types (see Section 5.1.1 for a discussion of the use of types vs. tokens).

4. Experimental setup

4.1 Corpora

We use two different types of data in this study: child-directed speech and written (adult-directed) text. Even though our aim here is not to model the child language acquisition process, the use of child-directed speech corpora makes the results more relevant to language acquisition research, providing evidence for the availability and consistency of the data available to children. On the other hand, we acknowledge that child-directed speech is expected to be marked in many aspects. To test whether the results obtained on child-directed speech are valid for larger set of naturally occurring language data, we also ran the same experiments on corpora of adult-directed written text.

Our child-directed corpora are obtained from the CHILDES (MacWhinney and Snow 1985) database. For Hungarian and Turkish we used all available child-directed utterances. The Hungarian corpus was collected by MacWhinney (1975), and it contains utterances directed to children between ages 1;5.2 (1 year 5 months and 2 days) and 2;10.22. The Turkish child-directed corpus (Slobin 1982) contains utterances directed to children between ages 2;0–4;4. For Dutch and English we choose sub-corpora that are similar in size, containing speech directed to younger (mostly pre-verbal) children.[5] Our English data comes from the corpus collected by Brent and Siskind (2001). We selected speech from 30 recording sessions with the youngest participants. The resulting sub-corpora included child-directed speech to 16 children in the age range between 0;8.27 and 0;9.28. For Dutch, we used recordings from the Groningen corpus (Bol 1995), and selected speech directed to the 25 youngest children in order to obtain a corpora of similar size. The resulting Dutch sub-corpora contained utterances directed to 5 children with age range 1;5.9–1;11.3.

Our adult-directed input comes from the Leipzig Corpora Collection (Quasthoff et al. 2006). For each language, we chose a segment of the corpora whose size is approximately 100,000 sentences. For all languages the corpora consist of randomly collected news text.

Table 3 presents the number of utterances, words and vowels in each sub-corpus we used in this study. Preprocessing (and also our method to count the vowels) is described in Section 4.2. Further discussion on the use of type or token counts can be found in Section 5.1.1.

[5] The choice of sub-corpora directed to younger children is motivated by the findings in the literature that children as young as 6-month old show sensitivity to vowel harmony if they are learning a VH language (van Kampen et al. 2008).

4.2 Preprocessing

First, we discard word tokens that contain non-alphabetic characters. We obtain the International Phonetic Alphabet (IPA) transcriptions with the help of an open source text-to-speech synthesizer (TTS).[6] Despite the fact that we found some inaccuracies of the TTS in our manual checks, we choose not to correct the output of the TTS, since the possible biases that may be introduced by the inaccuracies of TTS are unlikely to affect our investigation on VH.

After excluding monosyllabic words (containing only a single vowel), we remove all consonants from the transcribed text, so that only vowel sequences and word (and sentence/utterance) boundaries are retained. We consider diphthongs as single units at this step.

In this study, a vowel bigram is defined as two vowels with zero or more consonants in-between. Adjacent vowels as well as vowel bigrams that straddle word boundaries (since the scope of vowel harmony in most of the cases is the word)[7] are not used in vowel-bigram frequency calculations.

5. Finding vowel harmony

If a language exhibits vowel harmony, we expect the vowels that 'harmonize' to co-occur within the same word, while the vowels that do not harmonize tend not to co-occur within words. The co-occurrence of two (or more) vowels can be calculated from an appropriate corpus or lexicon. Crucially, however, vowel harmony does not predict co-occurrence of any two vowels but the vowels that harmonize according to one or more phonetic dimension. In non-VH languages we do not expect any particular pattern of co-occurrences of vowels that harmonize. We may observe some strong co-occurrence tendencies, but the important difference between VH and non-VH languages is that the co-occurrence tendencies in a VH language are expected to be related to the articulatory-phonetic features of the vowels.

Most computational studies of vowel harmony (e.g. Baker (2009); Goldsmith and Riggle (2012)) focus on the co-occurrence of vowels, typically fitting a model on vowel co-occurrence statistics in a corpus. The relation to the articulatory-phonetic features is established after the model fit. In this section, we will first follow this practice by analyzing the corpora of four languages by investigating co-occurrence of vowels in each corpora. After this analysis, we will provide an analysis that incorporates both the articulatory-phonetic features and the co-occurrence of vowels. Before presenting both types of analyses of the data on the four languages, we define the metric we use for measuring the co-occurrence of two vowels.

5.1 Pointwise mutual information

Like some of the previous studies (Goldsmith and Riggle 2012, Baker 2009), we use *pointwise mutual information* (PMI) as the metric for measuring the co-occurrence of two events, in this case two consecutive vowels with one or more intervening consonants. In this document, we use the term *vowel bigram* for two vowels with at least one consonant in between.[8] Hence the vowel bigram `a(C)*e` refers to an occurrence of vowel `a` followed by zero or more consonants and a vowel `e`[9] within the same word. The PMI of a vowel bigram `a(C)*e` is defined as

$$\mathrm{PMI}(\texttt{a(C)*e}) = \log_2 \frac{p(\texttt{a(C)*e})}{p(\texttt{a}_l)p(\texttt{e}_r)}$$

[6] We use the intermediate representations of eSpeak text-to-speech software available from http://espeak.sourceforge.net/ to convert text forms to IPA transcriptions.

[7] The particles that are affected by vowel harmony (see Section 3.2) are treated as separate words.

[8] We treat diphthongs as two separate vowels, where the initial part of the diphthong forms a vowel bigram with the vowel on its left and final part of the diphthong forms a vowel bigram with the vowel on its right. Diphthongs themselves are not considered a vowel bigram, since we assume that this process is different than the relations between distant vowel sequences we are interested in.

[9] where 'ae' is not a diphthong

where $p(\texttt{a(C)*e})$ is the probability of the vowel bigram `a(C)*e`: the probability of the vowel `a` followed by one or more consonants and an `e` within a word. The probabilities $p(\texttt{a}_l)$ and $p(\texttt{e}_r)$ are probabilities observing `a` and `e` as left (first) and the right (second) vowels in a vowel bigram, respectively.

The probabilities in our calculations of PMI for a vowel pair are their empirical probabilities (relative frequencies) as observed over the word types found in the corpus. We perform add-one smoothing to include those bigrams which are otherwise missing from the corpus for the visualizations presented in Section 5.2, while we exclude the unobserved vowel pairs from the models presented in Section 5.3.[10]

The intuition behind the PMI measure is that, if `a` and `e` occur independently, we expect the probability of the vowel bigram `a(C)*e` to be the same as the product of the individual probabilities of vowels in corresponding positions in a vowel bigram. This would result in a PMI score close to 0, indicating that the mutual information is 0 bits. On the other hand, if two vowels are more likely to occur in this configuration, the probability of a vowel bigram will be higher than the independent probabilities of the individual vowels, in which case we expect a positive PMI score. If a vowel bigram occurs markedly less than one would expect from the frequencies of the individual vowels, then the probability of the vowel bigram will be less than the product of the probabilities of the individual vowels and the PMI value will be negative.

5.1.1 Types vs. tokens

The question whether frequencies in the lexicon (types) or the frequencies in actual speech (tokens) are more important when investigating phonological phenomena is open (Baker 2009). One can argue that frequent and infrequent words are phonologically equally important, and use the types thereof. On the other hand, if the goal is to model the language, by explaining regularities in the data, it may be better to use token counts.

The results reported in this paper are primarily based on *tokens-over-lexicon* counts, that is, vowel pair tokens calculated over word types. The motivation behind this choice is that we are interested in a lexical phenomenon, the construction of words and how vowels are distributed in them.[11] Using PMI scores calculated on vowel bigram types disregards the more frequent pairs that are observed in the words of the language. On the other hand, calculating scores on vowel bigram tokens over all word tokens in the corpus causes frequent words to be weighted heavily, and since exceptions and irregularities are expected to happen more with the frequent words, it is better to reduce their dominance. We also run our experiments on token counts (over all corpus), which yielded similar results. We report results for both where appropriate.

5.2 Vowel co-occurrence

Given our definition of the co-occurrence measure above, we first visualize the co-occurrences of all vowel pairs in all four languages we study. Tables 4–7 present the vowel-pair PMI values for Hungarian, Turkish, Dutch and English respectively. We calculate PMI values for each vowel bigram, and present these values in a tabular format where vowels on rows and columns are sorted primarily on front–back dimension. Within front–back divisions, we also sort for roundedness first, and height next. The rows represent the first vowel in a bigram, and columns represent the second vowel. For example the cell with row label `a` and column label `e` shows the PMI value for the vowel bigram `a(C)*e`.

Since the vowels in Tables 4–7 are ordered based on the articulatory classes of vowels, for VH languages, we expect positive and negative values to cluster together, forming 'blocks' of similar PMI

[10] We ran our experiments both with and without smoothing, and found no significant differences in the results.

[11] Both VH languages we study are morphologically complex, and a lexicon consisting all inflected words is not tenable for these languages. However, VH is part of the word formation process for these languages, and word-types should still be a better domain to study vowel harmony.

	front					back			
	U			R		U		R	
	H	M		H	M	L		H	M
	i	e	ɛ	y	ø	a	ɑ	u	o
i	0.11	-0.46	0.15	-0.53	-0.90	0.48	-0.07	-0.49	-0.04
e	-0.03	1.00	1.26	0.77	-0.95	-0.94	-1.79	-1.93	-1.88
ɛ	-0.26	1.02	1.14	1.17	0.54	-1.78	-1.98	-0.57	-1.54
y	-0.11	0.92	0.58	1.99	2.72	-3.67	-4.69	-4.22	-3.20
ø	-0.33	0.88	0.80	1.04	2.67	-1.94	-3.84	-2.04	-2.69
a	0.02	-0.47	-3.33	-1.38	-3.08	0.30	0.80	0.15	0.59
ɑ	0.09	-2.18	-2.78	-3.95	-2.59	0.45	0.68	0.59	0.66
u	0.52	-3.40	-3.61	-3.45	-5.40	0.82	0.37	1.41	0.30
o	0.02	-2.10	-3.38	-5.27	-3.32	0.58	0.73	0.13	0.75

Table 4: PMI distribution of vowels in Hungarian. The vowels are grouped primarily as front–center–back, then as (U)nrounded and (R)unded, and finally as (H)igh, (M)id or (L)ow. The blue cells indicate positive PMI values while orange cells indicate negative PMI values. The darker the shade of the color the higher the absolute PMI value.

values. Although some regularities may be observed for individual vowels for non-VH languages, we do not expect the groups of vowels to follow a similar pattern.

The Hungarian data presented in Table 4 shows a clear separation between front and back vowels. We see higher PMI values at the upper left and lower right parts of the table, while the PMI values are low on lower left and upper right parts. The 'neutral' vowel i is clearly distinguishable since it does not show a clear co-occurrence preference with neither back nor front vowels. However we do not observe the same effect on the other neutral vowel ɛ. Although far from being as clear as the front–back harmony, a pattern due to roundedness harmony of front vowels is also observable.

Like for Hungarian, we also observe a front–back effect in Table 5 for Turkish. Besides the front–back co-occurrence tendency, we also see the roundedness harmony. This is more pronounced on front vowels, where the distribution of rounded and unrounded vowels are more balanced. Besides the typical harmony process, we also observe that after an unrounded vowel only another unrounded vowel can follow, while some of the rounded vowels can be followed by unrounded vowels.

For Dutch and English, presented in Table 4 and Table 7 respectively, we do not observe any pattern that would indicate existence of a vowel harmony process. One can observe co-occurrence tendencies of some individual vowels. However, none of these tendencies are linked to the shared acoustic properties of the individual vowels.

In summary, a careful look at the PMI scores between the individual vowels already sets the languages with and without vowel harmony apart. Furthermore, one can also get further insight into the type of vowel harmony in the language of study. Note, however, that the visualization we present so far leaves finding the patterns in the vowel pair distribution and relating these patterns to the acoustic features of the vowels to the experimenter. Furthermore, even though PMI is a principled metric of co-occurrence, we do not have any straightforward way of making inferences about each PMI value we found. In other words, we cannot distinguish between reliable and unreliable PMI estimates presented in the Tables 4–7. The next section will address some these issues.

5.3 Analyzing the vowel pair distribution and the phonetic features together

In the previous section, we presented an analysis of vowel bigram PMI values, and showed that, with careful ordering of the vowels in pair-PMI matrices, one can find indications of vowel harmony in a

	front							back					
	U				R			U		R			
	H		M		H	M		H	L	H		M	
	i	ɪ	e	ɛ	y	œ	ø	ɯ	a	u	ʊ	o	ɔ
i	1.05	1.38	0.73	0.50	-4.47	-0.27	-4.75	-5.17	-1.81	-5.67	-6.69	0.92	1.68
ɪ	1.10	1.53	0.90	1.51	-1.25	-0.37	-2.27	-4.28	-2.37	-3.19	-5.79	0.27	-2.95
e	1.69	1.61	1.03	1.20	-2.98	-0.36	-2.84	-7.85	-3.07	-5.76	-6.78	-2.59	-4.67
ɛ	1.30	1.21	1.33	1.68	-2.67	0.54	-2.37	-5.96	-2.42	-3.28	-3.89	-3.28	-5.37
y	-3.22	-3.29	0.76	0.62	3.77	0.40	3.38	-6.09	-2.85	-3.00	-2.70	0.17	1.71
œ	-1.80	-5.20	2.02	1.20	3.74	0.81	3.30	-5.68	-3.56	-3.59	-4.61	-1.00	-4.09
ø	-3.78	-4.59	1.03	1.41	2.75	1.43	4.25	-5.06	-2.94	-1.97	-2.41	-1.39	-3.47
ɯ	-6.47	-3.96	-4.78	-4.90	-4.47	-0.26	-4.75	1.36	0.52	-5.67	-6.69	0.92	1.60
a	-1.13	-2.23	-1.08	-2.06	-5.93	-0.14	-4.63	1.32	0.82	-2.60	-2.65	-0.95	-3.17
u	-4.12	-4.93	-3.27	-3.81	-2.12	-0.50	-1.66	-6.99	0.12	2.07	2.19	0.69	1.32
ʊ	-4.67	-2.68	-3.57	-2.11	-2.67	0.53	-2.95	-4.96	1.08	0.99	2.07	-2.28	-3.37
o	-3.25	-2.90	-3.15	-3.37	-0.67	0.95	-1.22	-5.54	0.14	2.32	2.21	1.89	-0.63
ɔ	-4.81	-5.04	-5.71	-3.24	-3.81	-0.61	-3.09	-4.51	0.43	2.53	2.29	-0.84	-3.19

Table 5: PMI distribution of vowels in Turkish. The vowels are grouped primarily as front–center–back, then as (U)nrounded and (R)unded, and finally as (H)igh, (M)id or (L)ow. The blue cells indicate positive PMI values while orange cells indicate negative PMI values. The darker the shade of the color the higher the absolute PMI value.

	front								center		back						
	U					R				?	U		R				
	H		M		L	H	M					L	H		M		L
	i	ɪ	ɛ	e	a	y	ø	œ	ɵ	ə	ʌ	ɑ	ʊ	u	o	ɔ	ɒ
i	0.91	0.12	-0.86	0.46	-1.51	0.49	0.63	-0.11	-1.52	0.26	-0.26	-0.11	-0.05	-1.02	0.27	0.28	-1.75
ɪ	-0.57	0.51	-0.11	-1.60	-1.58	-1.16	-0.59	-0.59	-0.19	0.72	0.26	-0.45	-0.53	-1.82	-1.28	-0.69	-2.23
ɛ	-0.47	-0.13	0.80	0.13	-1.39	-1.53	-0.55	0.68	-1.96	0.31	-0.70	-0.48	-0.49	0.54	-1.10	0.20	-2.18
e	-1.30	-1.02	-1.39	-0.53	1.92	-0.55	-1.44	-1.44	-2.85	-0.67	-1.59	-1.37	-1.38	-2.08	-1.86	-1.53	-0.07
a	0.95	0.01	0.49	0.70	-0.22	0.88	-0.55	-2.13	0.50	-0.71	0.52	0.36	-1.49	-0.69	0.25	-0.21	0.40
y	-0.07	-0.20	-0.61	-1.01	-1.24	2.37	1.27	-0.73	1.44	0.27	0.12	0.34	0.91	-0.38	-0.57	-0.56	-0.78
ø	-1.95	-0.56	-0.60	-1.51	-3.13	0.10	2.08	1.08	0.26	0.71	0.93	-1.65	2.73	0.44	-1.34	0.84	1.03
œ	-0.18	0.21	-0.83	0.26	-1.36	1.87	2.86	2.86	2.03	-2.61	2.70	0.12	4.50	1.21	0.43	1.02	2.80
ɵ	0.16	-1.03	-0.20	-1.57	-0.87	0.04	0.02	0.02	2.37	0.47	0.87	-2.72	1.66	-1.62	-0.40	-0.23	-0.03
ə	0.09	0.63	0.32	0.82	-0.73	-0.37	0.75	1.20	0.99	-1.24	0.60	0.53	-1.07	0.93	1.06	1.01	1.56
ʌ	-0.18	0.21	-0.83	0.26	-1.36	1.87	2.86	2.86	2.03	-2.61	2.70	0.12	4.50	1.21	0.43	1.02	2.80
ɑ	-0.23	0.21	0.33	-1.29	-1.19	-0.81	-0.56	-1.15	-0.97	0.38	-0.72	0.60	-0.51	-0.62	-0.33	-0.07	-1.20
ʊ	-0.28	-0.89	-0.13	-0.85	-2.46	-0.23	0.75	0.75	-0.07	0.04	0.60	0.34	2.39	-0.89	2.03	0.50	0.70
u	-1.18	-1.27	-0.05	-1.23	-2.36	-1.94	-0.95	0.05	-1.78	0.52	-1.11	-0.52	0.69	2.89	-1.06	-2.79	-1.01
o	-0.14	0.31	-0.62	-0.97	-1.32	-1.09	-0.69	-0.11	0.07	0.52	-0.84	-0.18	-0.05	-1.75	0.79	0.18	-1.74
ɔ	-0.75	-0.12	-0.13	-0.08	-1.31	-1.20	0.52	0.79	-1.04	0.41	0.05	0.46	-0.16	-0.64	0.10	0.54	-1.85
ɒ	-0.18	-1.37	-1.42	1.67	-1.94	0.28	1.27	1.27	0.44	0.13	1.12	0.53	2.91	-0.38	-0.16	-0.56	1.22

Table 6: PMI distribution of vowels in Dutch. The vowels are grouped primarily as front–center–back, then as (U)nrounded and (R)unded, and finally as (H)igh, (M)id or (L)ow. The blue cells indicate positive PMI values while orange cells indicate negative PMI values. The darker the shade of the color the higher the absolute PMI value.

	front					**center**			**back**					
	U						?		**U**		**R**			
	H		**M**		**L**	**M**		**L**	**M**	**L**	**H**		**M**	**L**
	i	ɪ	e	ɛ	a	ɜ	ə	ɐ	ʌ	ɑ	u	ʊ	ɔ	ɒ
i	0.76	0.01	-0.67	-1.12	-1.57	-0.20	0.24	1.08	-0.98	-0.10	0.69	-0.39	-0.33	-1.55
ɪ	-0.22	0.14	-0.27	-0.21	-0.07	-0.88	0.11	-3.18	0.30	-2.36	-1.57	-0.65	0.32	-0.01
e	1.05	-2.38	0.85	0.98	0.53	2.89	-2.01	2.18	0.53	2.00	1.20	2.71	1.77	1.55
ɛ	-1.07	0.06	0.14	0.59	-0.85	0.77	0.30	-0.52	-0.85	-1.71	-0.92	-1.00	0.38	-2.16
a	0.76	0.23	-3.25	-1.12	-2.57	-1.20	-0.11	0.89	0.24	1.70	0.11	-1.39	-0.75	-1.55
ɜ	0.03	0.12	0.41	1.28	-0.49	0.87	-0.72	0.16	-0.49	-0.03	-0.82	1.69	0.74	0.52
ə	0.57	-0.28	1.61	0.62	0.99	0.54	-1.05	-1.76	0.29	-0.36	0.43	-0.23	0.64	1.07
ɐ	1.25	-3.76	1.27	1.40	1.96	0.51	-0.94	0.80	0.15	2.61	1.41	0.33	-0.61	0.75
ʌ	-1.08	-0.05	-0.28	-0.56	1.05	-0.23	0.01	-0.94	-0.01	-1.13	0.40	0.58	-0.77	0.42
ɑ	-1.16	-0.79	-0.37	0.76	-1.68	0.68	0.09	3.14	-0.10	1.37	0.57	0.50	0.55	-0.67
u	-0.35	0.16	-0.14	-2.01	-1.45	-0.09	-0.42	0.20	1.13	-0.99	1.80	-0.28	-1.22	-1.44
ʊ	0.09	-0.12	-0.12	-1.57	-0.69	-0.66	0.56	-1.37	-0.69	-1.55	-0.76	-0.84	0.22	-0.42
ɔ	-0.58	-0.06	-0.79	0.93	-0.10	0.26	0.35	0.55	-2.10	-0.64	-1.43	0.07	-0.87	-0.09
ɒ	-1.29	0.39	-2.50	-0.04	-1.81	0.55	0.07	-1.16	-1.22	-1.35	-1.14	1.37	-1.57	0.79

Table 7: PMI distribution of vowels in English. The vowels are grouped primarily as front–center–back, then as (U)nrounded and (R)unded, and finally as (H)igh, (M)id or (L)ow. The blue cells indicate positive PMI values while orange cells indicate negative PMI values. The darker the shade of the color the higher the absolute PMI value.

corpus (or word list), and if the language exhibits vowel harmony, the type of vowel harmony in the language can also be investigated with the same method. In this section we extend this method to make use of the knowledge of vowel features together with their distribution with respect to other vowels. In a nutshell, we fit a standard linear model that predicts the co-occurrence of two vowels from their agreement on certain phonetic features.

As before, we calculate PMI scores (see Section 5.1) for every vowel pair. Each vowel bigram is classified as harmonic or non-harmonic with respect to three articulatory features (backness, height, roundedness). Since our corpora are transcribed in IPA, we simply use the prototypical articulatory features[12] of the typical segment matching each IPA symbol obtained during the transcription process. We code height and front–back features with three levels, high/mid/low and front/mid/back, and the roundedness feature with two levels (rounded/unrounded).

As PMI is a measure of co-occurrence of (or attraction between) vowel pairs, we can check if the agreement on the features listed above could predict this co-occurrence. Hence, the precise question we ask is whether the match in articulatory features of vowels predict their co-occurrence in a language. If the language under study is a VH-language, there should be a relation between the harmoniousness (match in articulatory features) and co-occurrence of the vowel pairs. Furthermore, we are also interested in investigating the interaction of different kinds of harmony. We can observe if, for instance, in a language there is a roundedness harmony, but only for front vowels.

Formally, we fit a general linear model where our response variable is the PMI scores of the vowel bigrams, and the predictors are three indicator variables that indicate whether the vowels in the bigram match in one of the three articulatory features.[13] For example, if we have the vowel bigram

[12] The features are extracted based on the well-known vowel chart (e.g. International Phonetic Association (1999) p.ix). We have only two exceptions. In Hungarian and Turkish the sound transcribed as [a] is coded as a back vowel. Even though this sound is phonetically a front (or mid-front) vowel, in these languages it is phonemically a back vowel contrasting with [e] in front–back dimension.

[13] Since the model predicts a single continuous value (PMI) from a set of categorical variables, our model is equivalent to factorial ANOVA. However, we will present and discuss our results outside the hypothesis-testing context, that ANOVA results are typically interpreted.

		CHILDES		Leipzig	
Language		R^2	adj-R^2	R^2	adj-R^2
Hungarian	types	0.48	0.43	0.58	0.54
	tokens	0.40	0.34	0.58	0.54
Turkish	types	0.45	0.42	0.50	0.48
	tokens	0.41	0.37	0.51	0.49
Dutch	types	0.16	0.13	0.06	0.03
	tokens	0.13	0.10	0.08	0.05
English	types	0.11	0.07	0.10	0.06
	tokens	0.19	0.15	0.13	0.10

Table 8: Fit of the models to data in all four languages measured by R-squared. We also report 'adjusted R-squared', which corrects for by-chance model fit caused by large number of model parameters. We report results for both CHILDES and Leipzig corpora, and both word types (lexicon) and word tokens (corpora) extracted from each corpora.

[`i(C)*ø`] (a front-high-unrounded vowel followed by one or more consonants and then a front-mid-rounded vowel) our indicator variables are 1, 0 and 0 for frontness, height and roundedness harmony, respectively. As well as the individual features, we also use the full set of interactions so that we can characterize sub-types of the vowel harmony process.

5.3.1 Results and discussion

Following earlier studies, e.g. Baker (2009) and Goldsmith and Riggle (2012), we first present the overall model fit. Previous studies presented the model fit by number of bits saved by making use of vowel bigram information in comparison to a baseline model where only the relation between consecutive phonemes were encoded. Intuitively, the unsupervised models presented in these studies would represent the data with fewer bits for a VH language if the vowel bigram distribution is included in the model. Note, however, that the relation between the model fit and the vowel harmony is indirect. The information provided by vowel bigrams is not necessarily due to vowel harmony. Any regularity that may be encoded by vowel bigrams can reduce the description length. Nevertheless, an improved fit due to vowel bigram information is an indication of existence of vowel harmony, although it is not a sufficient indication by itself.

In our model, the link between vowel co-occurrences and the harmony is more direct. The 'harmony' of vowels in a vowel bigram is the only predictor of their co-occurrence. As a result, models fit well to the data only if the language under consideration is a VH language or a language that consistently prefers disharmonious vowel sequences. Otherwise, we expect model fit to be poor. Since preference for 'disharmony' can be observed from the model parameters, we will follow others and interpret better model fit as an evidence for vowel harmony unless model parameters indicates disharmony. As a measure of model fit across different corpora, we use the well-known measure of coefficient of determination, R^2 (also known as η^2 in the context of ANOVA).

Table 8 presents multiple-R^2 measures for the model fit in all four languages. We present results for both child-directed speech corpora, and the larger written-language corpora from the Leipzig collection, using vowel bigrams over the word types, as well as the word tokens (see Section 5.1.1 for a discussion).

At first sight, Table 8 confirms our expectations. Articulatory classes of vowels are able to explain around half of the variation of the PMI scores for Hungarian and Turkish, and the variation explained remains markedly lower for English and Dutch. Furthermore, we observe that the model fit is better for the larger corpora for the VH-languages, suggesting that most of the frequent words are not exceptions to vowel harmony. The better fit in Hungarian data in comparison to Turkish data is at least partly due to the smaller number of vowels (once corresponding long and short vowels are collapsed) in Hungarian. Even if they had the same degree of VH compliance in both languages, we would expect the variation within vowels belonging to the same VH class to be higher for Turkish, as there would be more vowels belonging to one harmony class on average.[14] The differences between model fit over word types and tokens also meet our expectations. Although these differences seem to disappear for the larger corpora, the model fit scores over word types are better than word tokens.

The model fit for both non-VH languages in our study is lower than the VH languages. Nevertheless, it is interesting to see close to 20% of the variation in the vowel bigrams co-occurrences in child-directed speech is explained by the harmony (or disharmony) of the vowel sequences. We will discuss these somewhat large R^2 values together with the detailed investigation of the VH types below. The differences between child-directed and adult-directed corpora, on the other hand, seem to be reversed in non-VH languages. The model fit is worse for the larger corpora for both Dutch and English. The differences between word type and tokens are rather mixed, and considering their small magnitude, these differences may be due to chance effects.

As a first step into characterizing vowel harmony (or more precisely, the interaction of vowel co-occurrence and articulatory harmony of the vowels) in the languages we study, we present side-by side box plots of the PMI scores for each harmony class in Figure 2. Due to space concerns, we only present the child-directed speech data here; the plots we obtained on Leipzig corpora are similar.

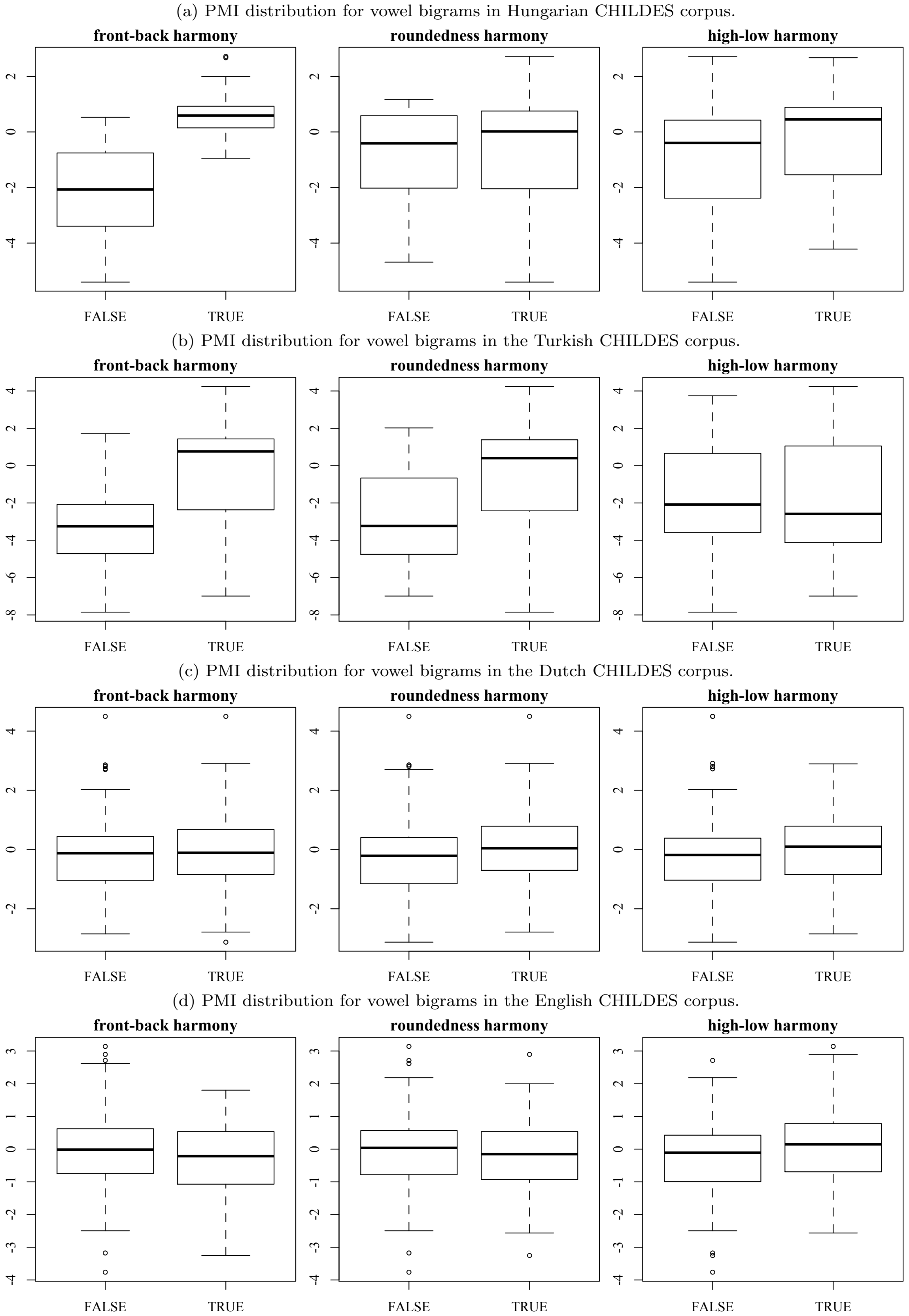
From the plots presented in Figure 2a–b, it is clear that there is a strong preference towards vowel bigrams that are harmonious with respect to their frontness. Turkish also shows a strong preference towards harmonious vowel bigrams with respect to their roundedness. The distributions of harmonious and non-harmonious vowel bigrams are similar with respect to height in both languages, as well as the roundedness in Hungarian. These plots already indicate that the harmony of the individual features affects the co-occurrence distributions of the vowel bigrams in a way that was expected from what we know about these languages.

For both non-VH languages in our study (Figure 2c–d) we do not observe any clear preference towards harmonious vowel bigrams. This finding is also clearly in accordance with the expectations. For English, we see a tendency towards vowel alternation (disharmony) with respect to frontness and roundedness. However, the magnitude of these differences seems to be small.

Although the box plots presented in Figure 2 allow us to learn about the distribution of harmonious and non-harmonious vowel bigrams in the data, they do not indicate whether these tendencies can be generalized outside the data set at hand. Furthermore, since we only inspected the primary dimensions of the articulatory features we are interested in, we cannot investigate the possible interactions between these features.

The estimated model parameters also serve as indications of vowel harmony. For VH languages we expect positive PMI estimates for vowel pairs that are harmonious, and negative PMI estimates otherwise. For non-VH languages, we do not expect any particular preference. Furthermore, we expect the estimated PMI values to be large and certain for the VH languages in comparison to the non-VH languages. We present the parameter estimates of models including all possible interactions in Table 9 for both child-directed speech and written corpora. The row labels indicate whether the harmony classes, i.e. the vowels bigrams that agree on one or more articulatory feature. For example, the rows labeled [$\pm front$] indicate the expected PMI values for a vowel pair that agree on front–back dimension but disagree on other two dimensions. The interpretation of interaction terms (the terms with a colon in between) is somewhat more difficult since they depend on the estimates of the other terms as well.

[14] In other words, if there is more variation to be explained due to the large number of vowels, it is more difficult to explain it. Hence, one expects a better measure of fit if the vowel inventory is smaller. As noted by an anonymous reviewer, this may result in a bias in interpretation of the model-fit measure. Although the sizes of vowel inventories are similar in our study, we note that the interpretation of R^2 needs caution.



(a) PMI distribution for vowel bigrams in Hungarian CHILDES corpus.

(b) PMI distribution for vowel bigrams in the Turkish CHILDES corpus.

(c) PMI distribution for vowel bigrams in the Dutch CHILDES corpus.

(d) PMI distribution for vowel bigrams in the English CHILDES corpus.

Figure 2: PMI distribution for vowel bigrams of Hungarian, Turkish, Dutch, and English, in the CHILDES corpora for each harmony class. The label TRUE indicates the bigrams that agree on the harmony dimension specified in the title of the plot.

	Hungarian	Turkish	Dutch	English
Intercept	-0.48 (0.33)	0.64 (0.46)	0.14 (0.10)	**0.17 (0.08)**
[±*front*]	**1.11 (0.36)**	-0.25 (0.52)	-0.23 (0.35)	-0.01 (0.31)
[±*high*]	-0.49 (0.52)	**-3.99 (1.63)**	0.06 (0.16)	0.24 (0.13)
[±*round*]	-0.10 (0.39)	**-2.05 (0.57)**	-0.25 (0.30)	-0.08 (0.17)
[±*front*] × [±*high*]	0.40 (0.65)	**4.83 (1.74)**	-0.38 (0.47)	**-1.25 (0.46)**
[±*front*] × [±*round*]	-0.31 (0.49)	**3.16 (0.63)**	**1.22 (0.47)**	-0.02 (0.37)
[±*round*] × [±*high*]	-1.86 (1.83)	2.10 (2.23)	-0.07 (0.51)	0.20 (0.39)
[±*front*] × [±*round*] × [±*high*]	2.68 (1.89)	-3.07 (2.32)	-0.05 (0.71)	0.92 (0.62)

(a) Parameter estimates on the CHILDES corpora.

	Hungarian	Turkish	Dutch	English
Intercept	-0.52 (0.28)	**-0.84 (0.36)**	0.07 (0.05)	0.04 (0.06)
[±*front*]	**1.23 (0.31)**	**1.12 (0.43)**	0.00 (0.14)	-0.02 (0.28)
[±*high*]	-0.59 (0.42)	-0.44 (0.60)	0.06 (0.08)	**0.32 (0.10)**
[±*round*]	-0.16 (0.33)	-0.24 (0.41)	-0.11 (0.12)	-0.03 (0.15)
[±*front*] × [±*high*]	0.60 (0.51)	0.93 (0.75)	**-0.40 (0.20)**	-0.48 (0.33)
[±*front*] × [±*round*]	-0.43 (0.42)	**1.12 (0.48)**	0.04 (0.19)	0.11 (0.32)
[±*round*] × [±*high*]	-1.01 (1.21)	-0.11 (1.19)	-0.09 (0.23)	**-0.86 (0.37)**
[±*front*] × [±*round*] × [±*high*]	1.66 (1.27)	-0.41 (1.28)	**0.62 (0.31)**	0.94 (0.50)

(b) Parameter estimates on the Leipzig corpora.

Table 9: Parameter estimates for the linear model including all interactions on the CHILDES (a) and Leipzig (b) corpora with all four languages in the study. The numbers between the parentheses after the parameter estimates are their standard errors. The parameters of indicator variables at the row labels indicate agreement of the vowel bigram in the specified articulatory feature. The estimates that would be considered statistically significant at a significance level of 0.05 are printed in boldface.

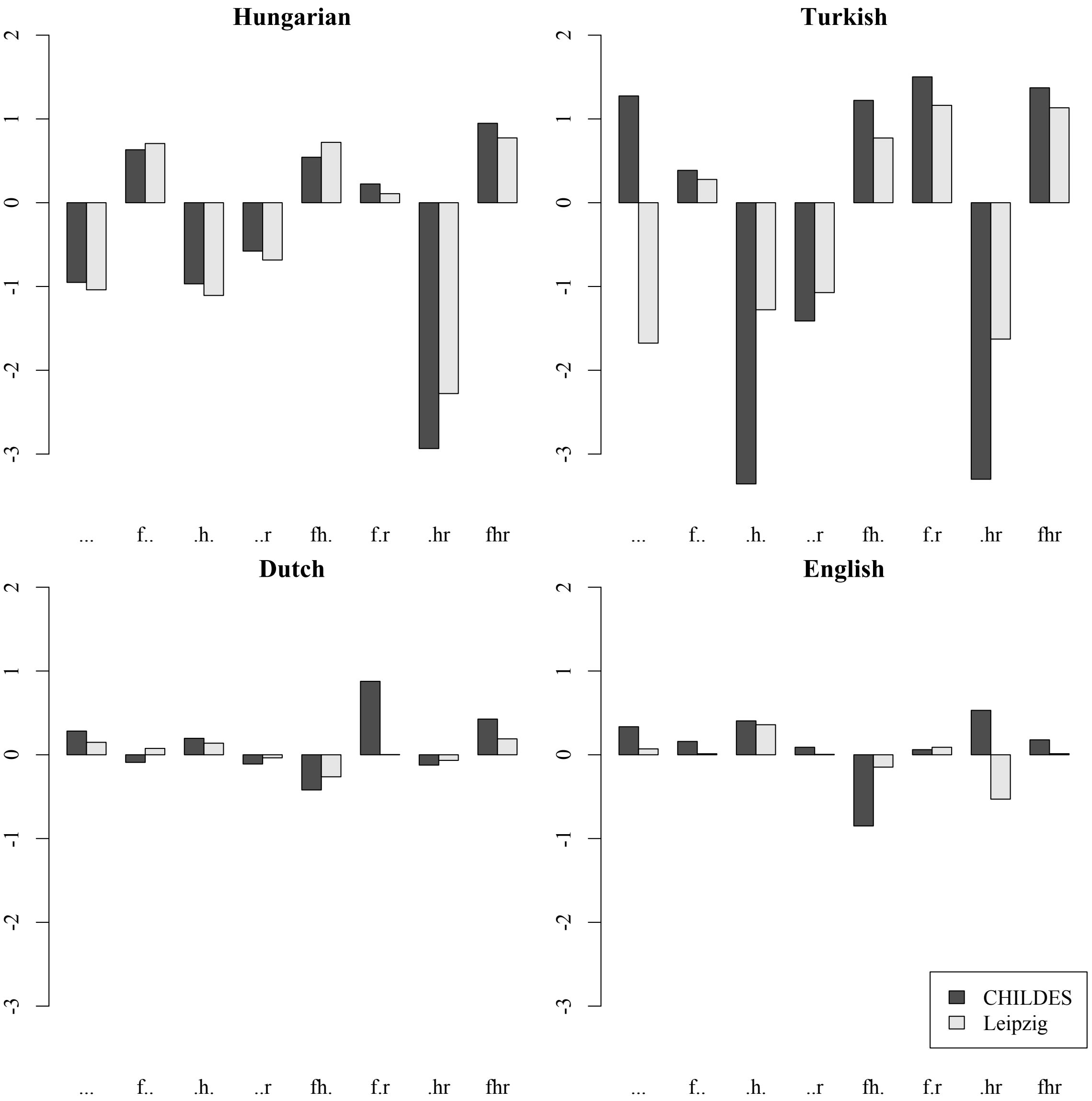


Figure 3: PMI estimates for all possible harmony classes in our model. The labels on the x-axes represent harmony of vowel bigram on front–back (f), height (h) and roundedness (r) features. If a particular label is missing (replaced with a dot) the vowel bigram is disharmonious with respect to this feature, e.g. 'f.r' means the bigram is harmonious with respect to front–back and roundedness features but not with respect to height.

Figure 3 intends to ease the interpretation of the data presented in Table 9. Instead of presenting the linear model parameters, Figure 3 presents the estimated PMI values for each possible VH case that can be represented by the model. In other words, we calculate expected PMI values (according to the fitted models) of all combinations of harmonious and disharmonious vowel bigrams in all three dimensions. Although it is difficult to make inferential statements from these calculated estimates, they provide a clearer picture with regards to the problem at hand.

As expected, Figure 3 indicates that absolute values of the PMI estimates are larger for the VH languages—both exhibit strong preference and dispreference of certain harmony cases. The estimated PMI values for Hungarian present a clear case for front–back harmony. All VH cases where the vowels are harmonious in the front–back dimension are preferred (estimated values are all positive), while all cases without front–back harmony are not preferred (estimated values are all negative). Turkish results are also similar, showing a clear dominance of front–back harmony. We also see the sign of roundedness harmony, since PMI estimates for bigrams that are harmonious both in front–back and roundedness dimension are higher than the case with only front–back harmony.

For both VH languages we observe consistent estimates for both corpora. The only exception seems to be the complete disharmony case (corresponds to the ‘intercept’ estimate of the linear model) for Turkish. Considering the uncertain estimate of the intercept parameter from the child-directed Turkish data, this is likely to be due to lack of enough numbers of completely disharmonious vowel bigrams in the child-directed speech corpus.

For the non-VH languages the estimates in both directions are close to zero. Although there are relatively large positive or negative PMI estimates in some cases these are mostly inconsistent in different corpora.

6. General discussion and conclusion

In this paper, we presented a computational/quantitative analysis of vowel harmony in four languages, two of which are known to exhibit vowel harmony. First, we have shown that in VH-languages harmonious vowels co-occur (as measured by pointwise mutual information between the vowels) more frequently than disharmonious vowels, while in non-VH languages one cannot find any preference towards harmonious vowels.

The approach in earlier studies (also in Section 5.2) have been to analyze only the distribution of vowels in a corpus. While this approach is attractive because of use of unsupervised methods, in such a study the relation of vowel co-occurrences and their harmony can only be established post-hoc. Furthermore, some non-harmonious co-occurrence preferences may increase the model fit, which has been used as an indication of vowel harmony in the input language. For example, in the Boltzmann framework of Goldsmith and Riggle (2012), it can happen, that a vowel preference due to a pattern observed in highly frequent words, or due to a tendency other than harmony may result in higher information gain in a bigram model.[15]

In this study we investigated the vowel harmony process using vowel co-occurrence information, together with the articulatory (or phonetic) features of the vowels. Our model predicts the co-occurrence of two vowels, as quantified by pointwise mutual information, from indicator variables representing harmoniousness of the vowel bigrams in multiple dimensions, front–back, high–low, rounded–unrounded. As expected, our model fits better to the data from VH languages compared to the data from non-VH languages. Besides better model fit, we show that the model predicts markedly higher co-occurrence for harmonious vowel pairs in VH languages. Our model's parameters can also be analyzed to gain further insights about the type of vowel harmony in the input language.

[15] For example, in Figure 2, and in some of the experiments we performed (but not presented here) we observed a tendency towards disharmonious vowel bigrams in English. This seems to be due to the fact that the non-stressed vowels are reduced in English and since stressed–non-stressed vowel pairs are common, the combination of a full (non-central) vowel and a schwa (central vowel) is also common.

Admittedly, our approach is linguistically informed and inherently supervised. However, we note that this is not necessarily a disadvantage. For the purposes of studying vowel harmony in a language quantitatively, we make use of the information available to linguists, and make the link between harmony and the co-occurrence explicitly. The results presented here show that one can characterize the vowel harmony of a language using a corpora and articulatory features of the vowels. Such models can be useful for investigating the properties of languages that are not studied as extensively as the languages studied in the current work.

From the language acquisition point of view[16] the cues presented—the co-occurrence statistics and articulatory-phonetic features of the vowels— are both available in the input. Although we do not make any claims about relevance of the current work to child language acquisition, our experiments show that these two sources of information correlate in the input language, and as a result may, at least in principle, facilitate learning the phonotactics of the input language.

Before closing, we list some points we consider as next steps for this work. Some of these points are direct extensions of this work to answer some questions which we believe to be interesting, some are related to shortcomings of the method used in the current study.

We addressed the question of how one can characterize the vowel harmony in a language. Another interesting question to ask is 'why some languages exhibit vowel harmony and some do not?' Generally, it would be interesting to see what phonetic, phonological, or even higher level, features correlate with the presence of VH in a language (e.g. size of vowel inventory, diphthongs, presence or lack of vowel reduction, rich or poor morphology, etc.).

Modeling vowel harmony using only vowel bigrams has the disadvantage of not revealing longer dependencies between the vowels in a sequence. A model paying more attention to the longer sequences may be able to discover more about the classes of the vowels in the language.

Additionally, in its current state our model does not characterize a vowel as 'neutral', as it only allows dichotomous categorization such as 'front' or 'back' classes for the backness feature. However, in Hungarian neutral vowels exist as well, that harmonize both with front and back vowels (for details see Section 3.1). In this study we labeled them according to their articulatory-acoustic features, but it would be interesting to see what happens if a subsequent model contained the neutral category as well.

Children learning VH-languages are known to be aware of this property of their language early on, and make use of it when learning other aspects of the languages, such as segmenting input speech into words. A future direction for this work is to model learning of vowel harmony, preferably within a general model of learning phonology of the language, in a way that would provide us insights about child language acquisition. With child language acquisition in mind, an unsupervised model making use of multiple cues (co-occurrence statistics and phonetic features) would naturally be a better choice than the supervised model we have presented here.

References

Baker, Adam C. (2009), Two statistical approaches to finding vowel harmony, *Technical Report TR-2009-03*, University of Chicago, Department of Computer Science.

Bellgrad, Matthew I. (1993), *Machine Learning of Temporal Sequences: Applications of the Effective Boltzman Machine*, PhD thesis, Univerisy of Western Australia.

Bol, Gerard W. (1995), Implicational scaling in child language acquisition: The order of production of Dutch verb constructions, *Papers from the Dutch-German Colloquium on Language Acquisition, Amsterdam Series in Child Language Development*, Vol. 3.

Brent, Michael R. and Jeffrey Mark Siskind (2001), The role of exposure to isolated words in early vocabulary development, *Cognition* **81**, pp. B33–B44.

[16] It is our intention to extend this work as a model of acquisition of vowel harmony.

Ellison, T Mark (1994), The iterative learning of phonological constraints, *Computational Linguistics.*

Goldsmith, John and Jason Riggle (2012), Information theoretic approaches to phonological structure: The case of Finnish vowel harmony, *Natural Language & Linguistic Theory* **30**, pp. 859–896, Chicago, IL, USA.

Hare, Mary (1990), The role of similarity in Hungarian vowel harmony: A connectionist account, *Connection Science* **2**, pp. 125–152.

International Phonetic Association (1999), *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, A Regents publication, Cambridge University Press.

Ketrez, F. Nihan (2013), Harmonic cues for speech segmentation: a cross-linguistic corpus study on child-directed speech, *Journal of Child Language* **FirstView**, pp. 1–23.

Kirchner, Robert (1993), Turkish vowel harmony and disharmony: An optimality theoretic account. 'Presented at Rutgers Optimality Workshop I (ROW-I).

Kiss, É. Katalin, Ferenc Kiefer, and Peter Siptar (2003), *Uj magyar nyelvtan*, Osiris, Budapest, Hungary.

MacWhinney, Brian (1975), Pragmatic patterns in child syntax, *Technical Report 198*, Carnegie Mellon University, Department of Psychology. http://repository.cmu.edu/psychology/198.

MacWhinney, Brian and Catherine Snow (1985), The child language data exchange system, *Journal of Child Language* **12** (2), pp. 271–269.

Manning, Christopher D. and Hinrich Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Quasthoff, Uwe, Matthias Richter, and Christian Biemann (2006), Corpus portal for search in monolingual corpora, *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp. 1799–1802.

Slobin, Dan I. (1982), Universal and particular in the acquisition of language, *in* Wanner, Eric and Lila R. Gleitman, editors, *Language Acquisition: The State of the Art*, Cambridge University Press, chapter 5, pp. 128–170.

Suomi, Kari, James M. McQueen, and Anne Cutler (1997), Vowel harmony and speech segmentation in Finnish, *Journal of Memory and Language* **36** (3), pp. 422–444.

van Kampen, Anja, Güliz Parmaksiz, Ruben van de Vijver, and Barbara Höhle (2008), Metrical and statistical cues for word segmentation: The use of vowel harmony and word stress as cues to word boundaries by 6- and 9month-old Turkish learners, *in* Gavarro, Anna and M. Joao Freitas, editors, *Language Acquisition and Development: Proceedings of GALA 2007*, pp. 313–324.