Morphophonemics and the lexicon: a case study from Turkish

Anne Pycha, Sharon Inkelas and Ronald Sprouse
University of California, Berkeley

To appear. M.J. Solé, P. Beddor, & M. Ohala (eds.) *Experimental Approaches to Phonology.* Oxford University Press.

Please cite published version; thank you!

1. Introduction

A large body of recent work has focused on statistical properties of the lexicon and their possible consequences for the grammar (e.g., Luce 1986, Content et al. 1990, Baayen 1993, Plag et al. 1999, Hay and Baayen 2002, Ernestus and Baayen 2003). The basic idea of this line of research is that every word can be meaningfully characterized by way of its relationship to other words in the same lexicon. Individual words are therefore not isolated entities; instead, their grammatical behavior can best be understood by considering lexicon variables such as frequency, neighborhood density, and cohort size.

For example, Rhodes (1992) claims that flapping in American English tends to occur in frequent words like *winter* but not in infrequent words like *banter*. Frequent words are produced with more gestural overlap and hence greater tendency for reduction, so the lexical statistics approach offers a plausible explanation for this gradient alternation. It is less clear what the approach offers for semi-regular alternations, where a majority of words undergo a categorical change but others do not. Consider stem-final alternations in Turkish nouns, in which coda /p, t, tʃ, k/ alternate with onset /b, d, dʒ, g/. Most nouns undergo the alternation (1a), but some resist it (1b).

(1)	a. kana**t**	wing	kana**d**-a	wing-DAT
	b. sana**t**	art	sana**t**-a	art-DAT

Traditional generative phonology partitions words like *kanat* and *sanat* into two categories: those that alternate and those that do not. This partition can be accomplished either with underlying representations or with indexed rules, but it is in any event category-based. The lexical statistics approach, on the other hand, claims that the difference between *kanat* and *sanat* can be attributed to how and where they are stored relative to other words in the lexicon, and that their alternation patterns should therefore correlate with lexicon variables.

These variables have been shown to impact word recognition and production processes in the laboratory (e.g., Wingfield 1968, Luce and Pisoni 1998, Marslen-Wilson and Welsh 1978, Gaskell and Marslen-Wilson 2002), but few or no studies exist which directly measure them, for Turkish or any other language. Instead, most studies have performed calculations using dictionaries, which are normative and conservative, or text corpora, which are compiled from disparate sources. Both text corpora and dictionaries represent the linguistic community as a whole, but neighborhood density and cohort size are ultimately facts about a single speaker's lexicon. The linguistic literature which speculates on the functional role of lexicon variables in natural language use (e.g., Wedel 2002) is thus getting ahead of itself, because the variables themselves have not really been measured.

In this paper, we employ a single-speaker corpus of Turkish to calculate lexicon variables, and to test for correlations with semi-regular alternations. This novel methodology allows us to model lexical storage more directly, without relying on text-based substitutes. Furthermore, it allows us to study the idiolectal, rather than the normative, application of semi-regular alternations. If the lexical statistics approach has teeth for semi-regular alternations as well as gradient ones, for example, we should find that words such as *kanat* and *sanat* reside in clusters of similar words that exhibit similar grammatical behavior, and these clusters should be characterizable by variables like neighborhood density, which have been claimed to correlate with applicability of gradient alternations. Our question is: do any of the variables predicting gradient alternation applicability also predict the applicability of semi-regular alternations?

Section 2 describes the Turkish alternations in more detail, and reviews a previous proposal, by Wedel (2002), for applying the lexical statistics approach to them. Section 3 describes the single-speaker corpus and overall methodology. The results of our examination of lexicon variables are presented in section 4 for word frequency, section 5 for cohorts, and section 6 for neighborhood density. Section 7 examines etymology. Section 8 concludes by reviewing the predictions of the lexical statistics approach, and comparing these predictions with those made by a generative account.

2. The problem

Our test case is the occurrence of stem-final alternations in Turkish nouns. These semi-regular alternations take two forms: velar deletion, in which /k/ and /g/ delete intervocalically when stem-final (2a), and plosive voicing alternations, in which coda /p, t, tʃ, k/ alternate with onset /b, d, dʒ, g/ (2b):[1]

(2) a. *Velar deletion*

bebe**k**	baby	bebe[Ø]-e	baby-DAT
arkeolo**g**	archeologist	arkeolo[Ø]-a	archeologist-DAT

b. *Plosive voicing alternations*

kana**t**	wing	kana**d**-a	wing-DAT
peroksi**t**	peroxide	peroksi**d**-e	peroxide-DAT
kita**p**	book	kita**b**-a	book-DAT
gyve**tʃ**	clay pot	gyve**dʒ**-e	clay pot-DAT
kepen**k**	shutter	kepen**g**-e	shutter-DAT

These alternations do not apply to verb stems.[2] Both patterns are productive in the sense of being applied to loans (e.g., *bifstek* [bifstek], *bifsteği* [bifstei] 'steak(-acc)'; *ofsayt* [ofsajt], *ofsaydı* [ofsajdɯ] 'offside(-acc)), but they are nonetheless semi-regular, rather than fully regular. As observed by Lewis (1967), most polysyllabic words undergo the alternations, but most monosyllabic words, plus a substantial subset of polysyllabic words, do not:

(3) a. Some polysyllabic words resist stem-final alternations

meto**d**	method	meto**d**-a	method-DAT
sana**t**	art	sana**t**-a	art-DAT

jour**t**	yogurt	jour**t**-a	yogurt-DAT
t͡ʃapu**t**	rag, patch	t͡ʃapu**t**-a	rag, patch-DAT
kre**p**	crepe	kre**p**-e	crepe-DAT
almana**k**	almanac	almana**k**-a	almanac-DAT
sinago**g**	synagogue	sinago**g**-a	synagogue-DAT

b. Most (C)VC words resist stem-final alternations

a**d**	name	a**d**-a	name-DAT
ko**d**	code	ko**d**-a	code-DAT
a**t**	horse	a**t**-a	horse-DAT
ke**p**	mortar board	ke**p**-e	mortar board-DAT
ha**dʒ**	pilgrimage	ha**dʒ**-a	pilgrimage-DAT
sa**tʃ**	hair	sa**tʃ**-a	hair-DAT
o**k**	arrow	o**k**-a	arrow-DAT
li**g**	league	li**g**-e	league-DAT

c. Some (C)VC words display alternations:

ta**t**	taste	ta**d**-a	taste-DAT
dʒe**p**	pocket	dʒe**b**-e	pocket-DAT
tʃo**k**	a lot	tʃo**[Ø]**-a	a lot-DAT

Inkelas and Orgun (1995) have offered a generative analysis of the Turkish facts, in which cyclicity and word minimality account for the special status of (C)VC stems, while lexical prespecification of consonant features accounts for the fact that within size categories, some nouns alternate while others do not. The prespecification analysis not only covers the relevant empirical ground, but also extends readily to other phonological problems such as geminate inalterability (Inkelas and Cho 1993) and nonderived environment blocking (Inkelas 2000).

Wedel (2002) has offered a lexical statistics account of the same facts, starting with the experimental finding, from other studies in the literature (e.g., Luce and Pisoni 1998), that high neighborhood density inhibits word recognition. Wedel suggests that if morphophonemic alternation also inhibits recognition, there may be a trading relationship in which the higher the neighborhood density of a word, the lower its probability of alternating and vice versa. According to Wedel, "If complexity within a lexical entry slows processing in any way (for example, if sub-entries compete with one another), alternation may be marked [i.e., less likely] in dense neighborhoods, where lexical access is already inefficient". Presumably the same argument should also hold for other lexicon variables that inhibit word recognition, such as low frequency (e.g., Wingfield 1968) or large cohort size (e.g., Marslen-Wilson and Welsh 1978, Gaskell and Marslen-Wilson 2002). Using lexicon variables calculated from a dictionary of Turkish, Wedel reports an inverse correlation between neighborhood density and alternation rate within (C)VC nouns.

The experimental literature actually presents a rather mixed view of the functional consequences of neighborhood density and surface alternations, casting doubt on the likely correctness of Wedel's hypothesis (on neighborhood density, see Luce and Pisoni 1998, Vitevitch and Sommers 2003; on surface alternations, see Otake et al. 1996, McLennan et al. 2003). Yet even if we could be sure that both neighborhood density and surface alternation had a consistent inhibitory effect in lexical access, Wedel's hypothesized relationship between lexicon variables and alternations cannot be meaningfully evaluated using statistics from a dictionary. Only

a single-speaker corpus provides an accurate model of lexical storage at the level of the individual speaker.

3. Methodology: TELL and a frequency corpus

The primary source of data for this study comes from the Turkish Electronic Living Lexicon (TELL), compiled at the University of California, Berkeley[3]. In addition, we use data on word frequency from a large text corpus developed by Kemal Oflazer at Sabancı University, Turkey.

3.1 TELL (Turkish Electronic Living Lexicon)

The Turkish Electronic Living Lexicon (TELL) is a searchable database of some 30,000 words in Turkish whose inflected forms have been elicited from a native speaker and transcribed phonologically (for a complete description of TELL, see Inkelas et al. 2000). The lexemes in TELL include roughly 25,000 headwords from the 2nd and 3rd editions of the Oxford Turkish-English dictionaries, along with 175 place names from an atlas of Istanbul and some 5,000 place names from a telephone area code directory of Turkey. A 63-year-old native speaker resident of Istanbul, who speaks standard Istanbul Turkish, was presented with the entire list, in random order. For each word in the list that the speaker knew – approximately 17,500 – he was asked to pronounce the word in isolation as well as in several other morphological contexts designed to reveal any morphophonemic alternations in the root. For nouns, words were elicited in the nominative case (no suffix), in the accusative case (vowel-initial suffix *-I*), in the first person predicative context (vowel-initial pre-stressing suffix *–(y)Im*), in the possessive case (vowel-initial suffix *–Im*) and with the derivational 'professional' suffix (consonant-initial suffix *-CI*).[4]

These pronunciations were transcribed and posted online on the TELL website, where users can search the list using regular expressions. In addition to the phonological transcriptions, both morphological and etymological information are provided. The morphological information shows the breakdown of complex stems, which were excluded from the current study and do not concern us here. The etymological information, provided for a large number of lexemes, gives the source language for the word. Thus, for example, a user interested in information about the Turkish word *kitap* 'book' could search TELL and turn up the following information:[5]

Table 1. Sample search of TELL

citation	**accusative**	**predicative**	**possessive**	**professional**	**etymology**
kitap	kitabı	kitabım	kitabım	kitapçı	Arabic

For all of the discussion that follows, any 'root' in Turkish may be either a stand-alone word (such as when *kitap* appears in the nominative) or the basis for further suffixation. For calculating lexicon statistics such as neighborhood density, we consider only those TELL entries in which the root is also a word (this includes the nominative *kitap,* but excludes the accusative *kitabı* and all other suffixed forms). Suffixed forms are used only to determine alternation rates.

3.2 Stem-final alternations: a snapshot from TELL

A search of all words in TELL, carefully hand-edited to eliminate errors as well as morphological compounds, yielded a total of 1,560 monomorphemic nouns that are potential candidates for stem-final alternations. Of these, approximately two-thirds end in plosives and are therefore potential undergoers of the semi-regular voicing

alternation. The remaining one-third end in velars and are therefore potential undergoers of the semi-regular velar alternation.

The actual alternation rates for the TELL speaker are shown in Table 2. These data confirm that stem-final alternations are in fact semi-regular for this speaker: they apply to most, but certainly not all, nouns in the lexicon. These data also show that alternation rates vary with word size, confirming Lewis's (1967) generalization that monosyllabic roots tend to resist these alternations.

Table 2. Overall noun count from TELL, and alternation rates

	% (C)VC nouns that alternate	% Longer nouns that alternate	Total
Voicing	17%	52%	1065
Velar deletion	8%	93%	495

Interestingly, the TELL data also reveal a significant effect of alternation type. While voicing alternation is three times as likely for longer roots than for (C)VC roots, velar deletion is over ten times as likely for longer roots.

3.3 Frequency corpus

To supplement the data on the single speaker represented in TELL, we used a text corpus compiled and morphologically analyzed by Kemal Oflazer, at Sabancı University in Istanbul, Turkey, for information about word frequency. The corpus consisted of some 12 million words drawn from newspapers, novels and other sources. The top and bottom ends of the word list, sorted by frequency, are illustrated in Table 3.

Table 3. Sample frequency counts

Turkish word	Gloss	Number of occurrences
ve	'and'	284663
bir	'one, a'	247294
bu	'this'	153903
da	'and, also'	94994
için	'for'	94642
de	'and, also'	94530
çok	'very'	55210
ile	'with'	51175
…		
mutsuzluktan		3
minyatürden		3
çıkıverdim		3
korunabilirsiniz		3
heyecanlandığımda		3

The most frequent item is a conjunction, *ve* 'and'; the least frequent item, *heyecanlandığımda*, is a highly morphologically complex word: *heyecan-lan-dığ-ım-da* 'excitement-VERBALIZER-PARTICIPLE-1.SG.POSSESSIVE-LOCATIVE'.

4. Frequency

The first variable that we considered as a possible correlate of the semi-regular pattern in Turkish is word frequency. Previous work has proposed direct correlations between frequency and grammatical behavior. Studies such as Rhodes (1992) on American English flapping and Bybee (2001) on coronal deletion have found that high frequency words undergo these alternations at higher rates than low frequency words do. Of course, both flapping and coronal deletion are gradient alternations that occur due to gestural overlap and are sensitive to speech rate. By contrast, the semi-regular Turkish alternations are not gradient, but categorical. That is, they apply categorically to a certain subset of Turkish nouns and are not affected by rate of speech. One question we ask in this section, then, is whether frequency has the same correlation with semi-regular, morphophonemic alternations as it does with gradient, phonetic alternations.

Frequency is somewhat different from the other variables examined in this study because it can be calculated in absolute terms, without reference to other items in the lexicon. The real explanatory value of frequency, however, can only be gauged in relative terms, by asking whether alternations tend to occur more often with high-frequency words than low-frequency words, or vice versa. If frequency has a role to play in the grammatical status of a given word, then, it would require a computation over the whole range of frequency counts in the lexicon.

To test for a frequency effect in Turkish, we took frequency counts from the text corpus described in section 3.3 and examined the correlation between mean frequency and alternation rate for nominal roots with lengths of one, two, and three syllables. The results are mixed.

In the case of roots which are candidates for velar deletion, alternators overall had a much higher mean text frequency than nonalternators.

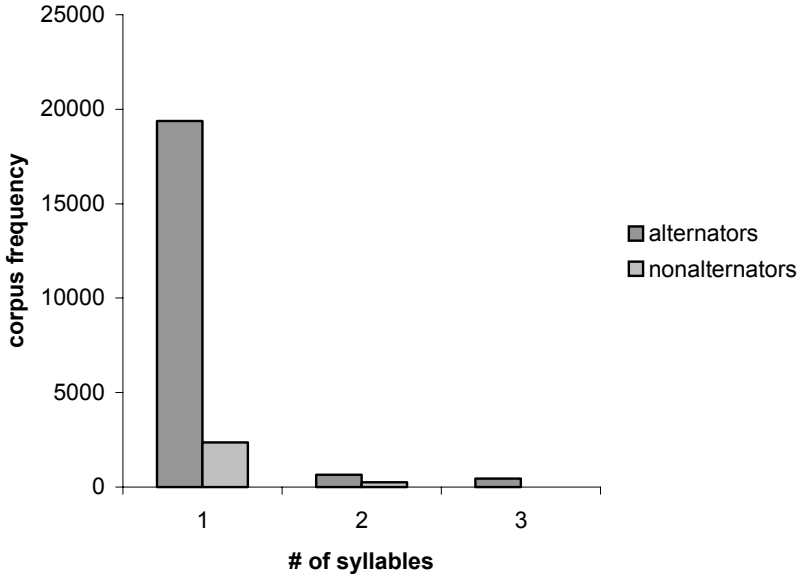

Fig. 1 Velar deletion: Cumulative root frequencies for alternating vs. nonalternating velar-final roots

For roots which are potential undergoers of voicing alternations, however, the results are the opposite: for one-syllable roots, the only category with roots of sufficiently high frequency to allow for comparisons, the class of alternating roots has a much lower mean frequency than the category of nonalternating roots.

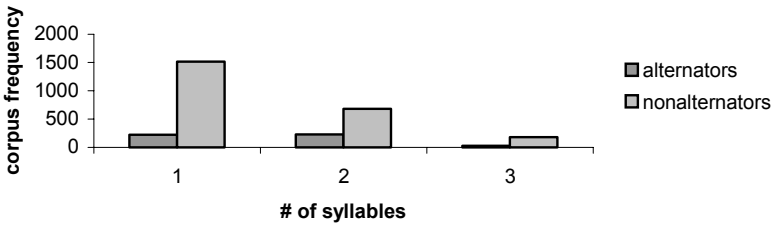

Fig. 2 Voicing alternation: Cumulative root frequencies for alternating vs. nonalternating nonvelar-final roots

Our findings for the velar alternation are thus in the same direction as that proposed by Rhodes (1992) and Bybee (2001) for gradient alternations in English: more frequent words are more likely to alternate. Our findings for the voicing alternation, however, are in the opposite direction: less frequent words are more likely to alternate. If frequency is to have any explanatory value for Turkish, then, its effects must be teased apart in a way that can account for these divergent results. One natural question to ask is whether the velar alternation has more in common with gradient, phonetic patterns (such as flapping) than the voicing alternation does, but in fact the opposite would appear to be true.

5. Neighborhood density

The next variable that we examined is neighborhood density, using standard measures from the literature (Luce and Pisoni 1998). For any given word, its 'neighbors' are those words which are most phonologically similar to it, and its 'neighborhood density' is the number of words that differ from it by the addition, deletion, or featural change of a single segment. For example, the Turkish word *amut* 'a perpendicular' has neighbors that differ by the insertion of one segment (***m***amut 'mammoth'), the deletion of one segment (*mut* 'happiness'), and the change of one segment (*a**v**ut* 'to soothe').

As discussed in Section 2, Wedel (2002) has proposed a direct correlation between neighborhood density and Turkish voicing and velar alternations. Using dictionary data, Wedel found that neighborhood density decreases with word length, a fact that is presumably true cross-linguistically. Specific to the Turkish alternation question, however, Wedel also found that within the class of (C)VC nouns, there was an inverse correlation between neighborhood density and alternation rate; the non-alternators had neighborhood densities that were approximately 40% higher than those of the alternators.

Wedel used a print dictionary of Turkish (the Redhouse Turkish-English dictionary; Alkım 1968) both as the source of information on individual word alternations and to calculate neighborhood density. The fact that a print dictionary includes many words that an average speaker might not know skews the computation

of neighborhood density to a large degree. A print dictionary also acts as a normative, rather than descriptive, source of information about alternations.

A single-speaker corpus, however, permits us to study only those words known to that speaker, and an idiolect-specific pattern of alternation. We therefore revisit the hypothesized correlation between neighborhood density and alternation rate, using TELL as the source for both neighborhood density calculations and word alternations.

5.1 Neighborhood density with a single-speaker corpus

Using the same definition of neighborhood employed in Wedel's study, we tested whether alternating nouns and nonalternating nouns had different mean neighborhood densities. To begin, we produced a list of neighbors for each monomorphemic noun in TELL; counting these neighbors produces the 'neighborhood density' variable. Below are examples of some of the nouns in the database and their neighbors. As discussed in section 2, only nouns can potentially undergo the alternations in question, but both nouns and verbs are included as potential neighbors for nouns.

Table 4. Some near-minimal pairs for neighborhood density

Root	**Alternates? (accusative shown)**	**Neighbors**	**Neighborhood density**
aɯɪt	Yes; aɯ**d**-ɯ	azɯt, ait, acɯt, ast, ant, kɯt, aɯn, aɯz, at, aɯl, anɯt, zɯt, art, baɯt, aft, alt, aut, arɯt, akɯt, aɯr	20
ait	No; ai**t**-i	vait, bit, ast, ant, git, aɯt, at, dʒit, asit, it, alt, art, aft, fit, eit, aut, akit, sit, zait, ahit, mit	21
amut	Yes; amu**d**-u	mamut, anut, umut, aut, mut, armut, hamut, avut	8
anut	No; anu**t**-u	amut, aut, unut, ant, anot, anɯt, angut, avut	8
armut	Yes; armu**d**-u	amut	1
angut	No; angu**t**-u	anut	1

Our first finding is that mean neighborhood density is inversely correlated with word length:

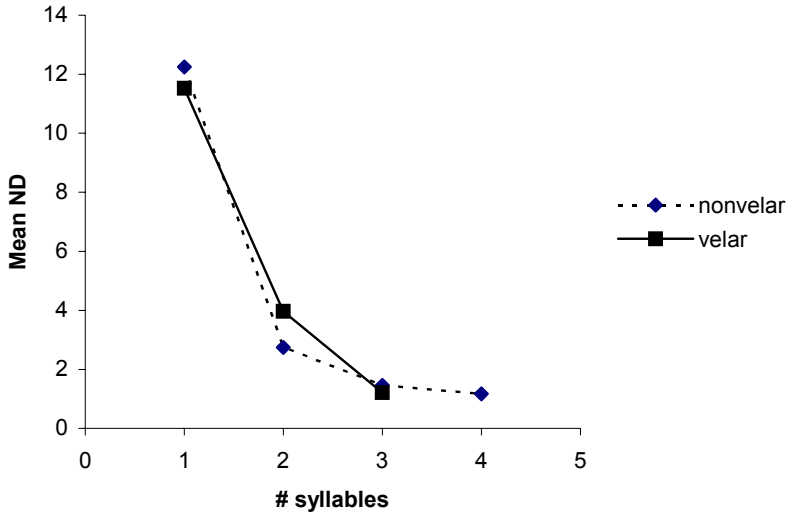

Fig. 3 Mean neighborhood density as a function of root length: Mean neighborhood density declines with root length

This finding is similar to what Wedel found using data from the Redhouse dictionary.

Unlike the dictionary study, however, we found no correlation between mean neighborhood density and alternation rate. As seen in the charts in Fig. 4 for voicing alternations and Fig. 5 for velar deletion, while mean neighborhood density varies according to word length, it is virtually identical for alternators and nonalternators within a given word-size category, as measured by syllable count (note the almost complete overlap in Fig. 4 between the solid and dashed lines).

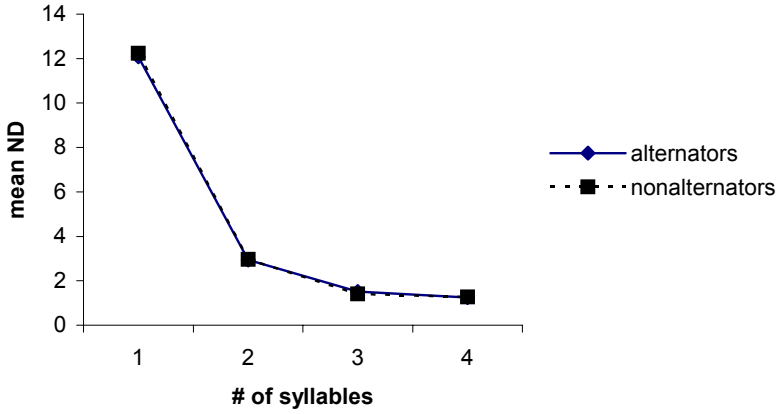

Fig. 4 Voicing alternation: Mean neighborhood density of alternating vs. nonalternating plosive-final roots

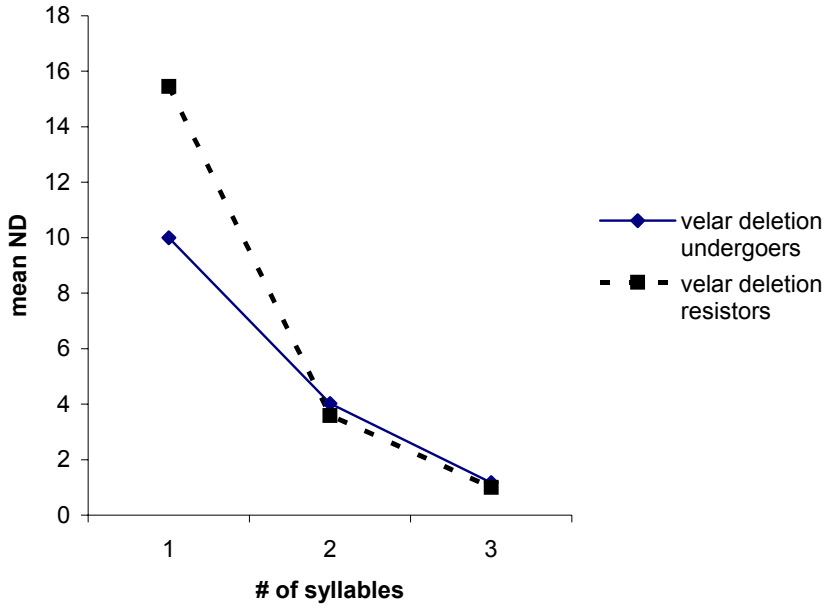

Fig. 5 Velar deletion: Mean neighborhood density of alternating vs. nonalternating velar-final roots

The only size category for which we observed mean neighborhood density to be a factor in alternation rates is monosyllabic velar-final roots. Mean neighborhood density for nonalternators is higher than that for alternators (15 vs. 10), but the difference is not significant.

Neighborhood density can be defined in several ways, however. If the speaker/listener performs computations that refer only to highly similar words in the lexicon, a narrower definition of 'neighbor' could yield different results. In particular, since the Turkish alternations affect stem-final segments, we can hypothesize that neighborhood densities calculated according to changes in the final segment (as opposed to the addition, deletion, or change of a segment anywhere in the word) are more likely to correlate with alternation rates.

We tested this hypothesis using neighbors that differed only in the final segment. For example, under this definition, the word [aɯt] has just four neighbors (instead of 20): [aɯn], [aɯz], [aɯl], and [aɯr]. We also tested the hypothesis using neighbors that differed only in the voicing of the final segment. For example, the word [at] has the neighbor [ad], but the word [aɯt] has no neighbors (*[aɯd]). Changing the definition of neighbor had no effect on our findings. The results were the same.

In summary, then, there is no evidence that neighborhood density explains the semi-regular pattern of alternation in Turkish; there is only evidence that mean neighborhood density is correlated with word size. We know that alternation rate is also correlated with size, but in a different way. Mean neighborhood density declines abruptly from monosyllabic to polysyllabic words (see Figs. 4 and 5), but within polysyllabic words, it declines gradually with increasing size. By contrast, as shown

in Fig. 6, alternation rates increase abruptly at the one-syllable mark but hold relatively steady across 2-, 3- and 4-syllable roots:[6]

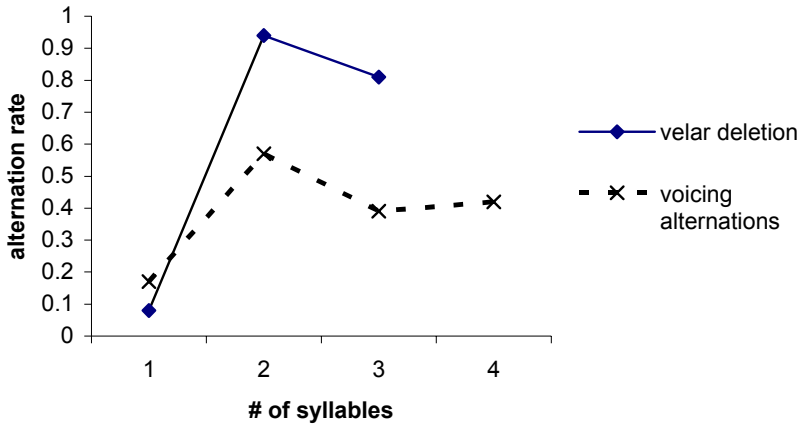

Fig. 6 Alternation rate as a function of number of syllables in the root

This categorical difference between 1-syllable roots and all others is also recognized by Wedel (2002), who ultimately concludes that the special behavior of one-syllable roots cannot be explained synchronically by mean neighborhood density, although it may be, according to Wedel, the result of grammaticalization of what once was a statistical lexical effect of mean neighborhood density on alternation.

5.2 Frequency-weighted neighborhood density

Thus far we have found that neither word frequency nor neighborhood density correlates in a consistent way with alternation rate (word frequency makes opposite predictions for velar deletion vs. voicing alternations, while neighborhood density makes no predictions at all). It is conceivable, however, that the speaker/listener performs a computation over the entire lexicon using multiple variables at once. A composite variable, then, might reveal correlations with alternation rate that single variables do not.

A composite variable can also address one problem with interpreting the frequency results presented earlier: the impact of a single, highly frequent word can be enormous, obscuring what might be a more general pattern. For example, the word *çok* [tʃok]'very', a monosyllabic word that undergoes the velar deletion alternation (*çoğ-u* [tʃo[Ø].u] 'very-ACC'), occurs 55,210 times in the corpus, a number which may well be responsible for the spike in the chart in Fig. 1.

We therefore calculated a composite lexicon variable called 'neighborhood frequency'. This is the sum of the corpus frequencies of the neighbors of each root. Our example minimal pair for neighborhood density ([aɯɪt], [ait]) is a near-minimal pair for neighborhood frequency, as illustrated below:

Table 5. A near-minimal pair for neighborhood frequency

Word	Alternates? (accusative shown)	Number of neighbors	Neighborhood frequency
aut	Yes; *au**d**-u*	20	113,091
ait	No; *ai**t**-i*	21	106,297

Our results indicate that in general, there is no correlation between neighborhood frequency and alternation rates. For the voicing alternation, mean neighborhood frequencies are almost identical for both alternators and nonalternators.

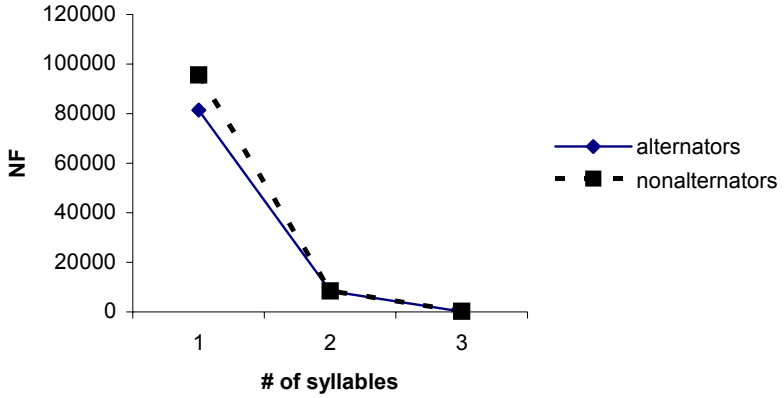

Fig. 7 Voicing alternation: Neighborhood frequency for alternators vs. nonalternators

For the velar alternation, two- and three-syllable roots also show no correlation. For monosyllabic roots, though, higher neighborhood frequencies are correlated with the nonalternating pattern.

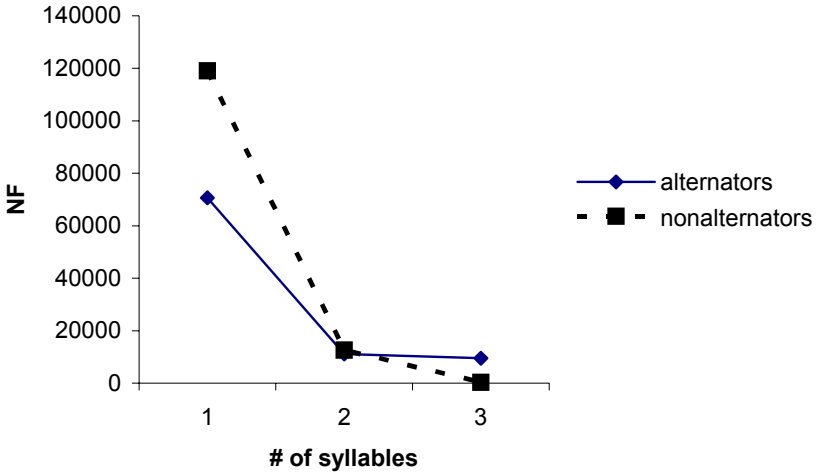

Fig. 8 Velar alternation: Neighborhood frequency for alternators vs. nonalternators

Recall that word frequency exhibited very different relationships to voicing and velar deletion. Neighborhood frequency, on the other hand, shows a more consistent pattern across the two alternation types, at least for words that are two or more syllables in length. And for both alternations, neighborhood frequency is highest for one-syllable roots and roughly even for roots of two and three syllables.

In summary, neighborhood frequency does not in general correlate with alternation rate. Instead, what we are seeing is a large difference between monosyllabic and larger roots, with no difference in mean neighborhood frequency between alternators and nonalternators in longer roots.

6. Cohorts

The final lexicon variable that we tested is cohort size (and its related variable, uniqueness point) (Marslen-Wilson and Welsh 1978, Luce 1986, Gaskell and Marslen-Wilson 2002). Starting from the initial segment of a word, the 'uniqueness point' is that segment at which the word becomes uniquely distinguishable from all other words in the lexicon. Some words have no uniqueness point at all. In the current study, the 'cohort' of a word is that set of words sharing the initial substring ending one segment before the uniqueness point. Thus, for example, in TELL the root *sokak* 'street' has a uniqueness point at the 4th segment, [*a*]. No other word in the lexicon begins with the substring *soka-*, so *sokak* is uniquely identifiable at [*a*]. However, five words (including *sokak* itself) begin with the substring *sok-*; this is the cohort of *sokak:*

(4) *sokak* street
Uniqueness point at 4th segment (*soka-*)
Cohort size = 5 (*sokul, sokuş, sokum, sokak, sok*)

Like neighborhood density, cohort size is a way of measuring a word's phonological similarity to the rest of the lexicon. But while neighborhood density is by definition restricted to measuring words of a similar length (neighbors can differ

by one segment at most), cohort size is not so restricted. Using cohort size as a lexicon variable, then, allows us to test whether phonologically similar words of any length are relevant for the speaker/listener's hypothesized computations over the lexicon.

Using TELL, we examined cohort sizes as well as uniqueness points for (C)VC roots and longer (>CVC) roots, looking separately at voicing alternations and velar deletion. We compared cohorts of small (2-3 words), medium (4-19 words), and large (20 and up) sizes, as well as uniqueness points that are early (before the last segment), late (at the last segment), and nonexistent (roots with no uniqueness point). The findings were negative: in no case did we find that cohort size or uniqueness point correlated with alternation rates. Rather, alternation rates were remarkably stable across all cohort and uniqueness point categories, with one exception.

Consider, for example, the figures for voicing alternation in longer roots, shown in Table 6.

Table 6. Voicing alternation in roots longer than CVC

	Cohort size		
	Small	Medium	Large
	2-3	4-19	>19
Number of alternators	242	174	49
Number of nonalternators	505	333	81
% Alternators	32%	34%	38%

The small differences in alternation rates across the cohort size groups is not significant. Figures for velar deletion are similar.

For uniqueness point, we also found essentially no difference, and certainly no statistically significant difference, except in one case: voicing alternations in roots longer than CVC.

Table 7. Voicing Alternation in roots longer than CVC

	Uniqueness point		
	Early	Late	None
Number of alternators	244	196	25
Number of nonalternators	662	189	68
% Alternators	27%	51%	27%

In this case, the difference between roots with a late uniqueness point and all other roots (early or no uniqueness point) is significant. Precisely in roots where the final consonant served to uniquely disambiguate the root, more alternation was found than in all other roots.

7. Etymology

A simple hypothesis that immediately springs to mind but can almost as easily be dismissed is that whether or not a given noun alternates is predictable from its etymological status. If voicing alternations and velar deletion are frozen morphophonemic alternations, perhaps it is simply the case that native nouns exhibit them and loans do not. Data to test this conjecture are available in TELL, which supplies etymological information for 7,917 lexemes. Of these, the majority (6,014, or

76%) are identified as non-native. This proportion is undoubtedly exaggerated due to the fact that the etymological information in TELL came in part from dictionaries of loanwords, and is therefore biased in the direction of non-native words. Nonetheless, Turkish is noted for having borrowed heavily from a variety of languages – for example, Arabic, Persian, French, Italian, Greek – and thus it is not surprising that TELL contains so many loanwords.

The chart below illustrates graphically the proportion of 1, 2 and 3-syllable nouns in TELL which are identified as native, identified as nonnative, and not identified either way:

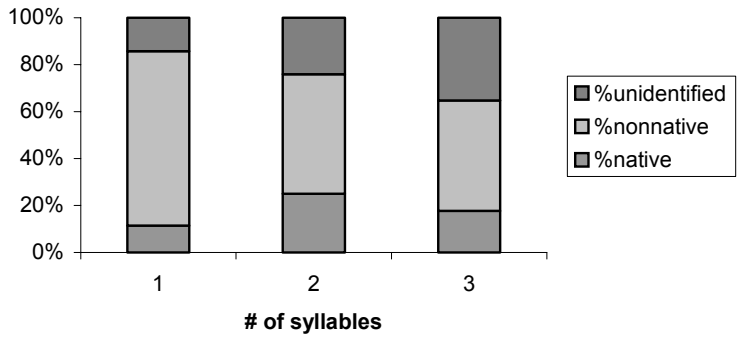

Fig. 9 Etymological sources of nouns by size

The statistical question at issue is whether native nouns differ from nonnative nouns in their overall rates of alternation. The figures in Table 8 pose this question for roots ending in *-t* or *-d*, i.e., coronal-final roots that are potential participants in the voicing alternation. As seen, the alternation rates of native vs. nonnative words are similar. While the number of nouns identified in the TELL database as native (which likely underestimates the actual representation of native roots in the lexicon) is perhaps too small to generalize from (12 only), the number of nonnative nouns is quite large (412), and it is clear that for these nouns, alternation is quite a common pattern.

Table 8. No pattern of alternation based on native status

	Polysyllabic coronal-final nouns
Native	50% alternation rate (n = 12)
Nonnative	42% alternation rate (n = 412)

Many of the words in the alternating nonnative category are quite familiar items; the list includes *kili**t**, kili**d**-i* 'lock(-ACC)', from Greek; *armu**t**, armu**d**-u* 'pear(-ACC)', from Farsi; and *orki**t**, orki**d**-i* 'orchid(-ACC)', probably from French

In summary, etymological origin appears not to be a factor in overall alternation rates. As has often been remarked on in descriptions of Turkish, a large portion of the lexicon is non-native to begin with, and nonnative words, old and new,

are readily assimilated into patterns of morphophonological alternation, rather than resisting alternations in numbers large enough to account for the figures in Table 2.

8. Conclusions

Thus far we have found no lexicon variable that predicts or explains why (C)VC roots in Turkish alternate less than longer roots and why voicing alternations are less consistent than velar deletion within words of a given size. When a single-speaker corpus is employed, an appeal to one word's relationship with other words in the lexicon – in the form of neighborhood density, neighborhood frequency, or cohort size – does not predict whether a word is more or less likely to undergo alternation. (By contrast, recall from section 4 that we did find a positive correlation between raw word frequency, as calculated from a large text corpus, and velar alternation rate, though there was an inverse correlation in that same corpus between raw frequency and voicing alternation rate). Our single-speaker study thus finds no support for Wedel's (2002) hypothesis that roots with larger neighborhood densities (and presumably, lesser frequencies and larger cohorts) will be less amenable to undergoing phonological alternations.

For existing roots in Turkish, there is simply no way to predict whether a given root will alternate other than to hear the accusative (or other relevant suffixed) form. The more insightful analysis for Turkish velar deletion and voicing alternations thus appears to be the generative one, in which the underlying representation of each root contains the information necessary to predict its grammatical behavior (Inkelas and Orgun 1995). Words undergo, or resist, morphophonemic alternation in a manner unrelated to the noun's relationship to other lexical items.

References

Alkım, U. B. (1968). *Redhouse Yeni Türkçe-İngilizce Sözlük [New Redhouse Turkish-English Dictionary],* Istanbul, Redhouse Yayınevi.

Baayen, H. (1993). On frequency, transparency and productivity. *Yearbook of Morphology 1992***,** 181-208.

Bybee, J. (2001). *Phonology and language use,* Cambridge, Cambridge University Press.

Content, A., Mousty, P. and Radeau, M. (1990). BRULEX: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique,* 90**,** 551-566.

Ernestus, M. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79**,** 5-38.

Gaskell, M. G. and Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology* 45**,** 220-266.

Hay, J. and Baayen, H. (2002). Parsing and productivity. *Yearbook of Morphology 2001***,** 203-235.

Inkelas, S. (1995). The consequences of optimization for underspecification. In Beckman, J. (ed.) *Proceedings of the Northeastern Linguistics Society 25:* 287-302. Amherst, MA, Graduate Linguistic Student Association.

Inkelas, S. (2000). Phonotactic blocking through structural immunity. In Stiebels, B. and Wunderlich, D. (eds.) *Lexicon in focus.* Berlin, Akademie Verlag, 7-40.

Inkelas, S. and Cho, Y.-M. Y. (1993). Inalterability as prespecification. *Language,* 69**,** 529-74.

Inkelas, S., Küntay, A., Orgun, O. and Sprouse, R. (2000). Turkish Electronic Living Lexicon (TELL). *Turkic Languages,* 4**,** 253-75.
Inkelas, S. and Orgun, C. O. (1995). Level ordering and economy in the lexical phonology of Turkish. *Language,* 71**,** 763-793.
Lewis, G. (1967). *Turkish grammar,* Oxford, Oxford University Press.
Luce, P. A. (1986). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics* 39**,** 155-158.
Luce, P. A. and Pisoni, D. B. (1998). Recognising spoken words: the neighborhood activation model. *Ear & Hearing,* 19**,** 1-36.
Marslen-Wilson, W. D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology,* 10**,** 29-63.
Mclennan, C. T., Luce, P. A. and Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 29**,** 539-553.
Orgun, C. O. (1995). Correspondence and identity constraints in two-level Optimality Theory. In Camacho, J. (ed.) *Proceedings of the 14th West Coast Conference on Formal Linguistics.* Stanford, Stanford Linguistics Association, 399-413.
Otake, T., Yoneyama, K., Cutler, A. and Lugt, A. V. D. (1996). The representation of Japanese moraic nasals. *Journal of the Acoustical Society of America,* 100**,** 3831-3842.
Plag, I., Dalton-Puffer, C. and Baayen, H. (1999). Morphological productivity across speech and writing. *English Language and Linguistics,* 3**,** 209-228.
Rhodes, R. (1992). Flapping in American English. In Dressler, W., Prinzhorn, M. and Rennison, J. R. (eds.) *Phonologica 1992.* Torino, Rosenberg & Sellier, 217-232.
Sezer, E. (1981). The k/Ø alternation in Turkish. In Clements, G. N. (ed.) *Harvard Studies in Phonology.* Bloomington, Indiana University Linguistics Club, 354-382.
Vitevitch, M. and Sommers, M. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition,* 31**,** 491-504.
Wedel, A. (2002). Phonological alternation, lexical neighborhood density and markedness in processing. Unpublished presentation, Eighth Conference on Laboratory Phonology.
Wingfield, A. (1968). Effects of frequency on identification and naming of objects. *American Journal of Psychology,* 81**,** 226-234.
Zimmer, K. and Abbott, B. (1978). The k/Ø alternation in Turkish; some experimental evidence for its productivity. *Journal of Psycholinguistic Research,* 7**,** 35-46.

Notes.

[1] Velar deletion is systematically inhibited following a phonemically long vowel, as in words like [meraːk-ɯ] *merakı* 'curious-acc'; see Sezer 1981 On plosive voicing alternations and velar deletion, see Lewis 1967, Zimmer and Abbott 1978, Inkelas 1995, Inkelas and Orgun 1995, Orgun 1995, Inkelas 2000, among others. The data in (2) are presented in IPA. In the dialect we describe, velar deletion is absolute; in other more conservative dialects, velars fricate or glide in the same environment.

[2] Velar deletion does apply, however, to velar-final suffixes, e.g., the future: *al-a͡ʤak-sin* 'take-FUT-2sg = you will take', *al-a͡ʤa[Ø]-ız* 'take-FUT-1PL.SUBJ = we will take'.

[3] TELL is available online at http://linguistics.berkeley.edu/TELL.

[4] Following standard convention, uppercase letters in the representations of affixes refer to segments which alternate, predictably, for a particular feature; in the case of vowels, the harmonic features [back] and [round], and in the case of consonants, [voice].

[5] Transcriptions are presented, in this paper, in Turkish orthography or in IPA, as appropriate. This example uses orthography. In the TELL database itself, an ascii transcription is used instead, with a code allowing users to translate to orthography and/or IPA.

[6] There were no 4-syllable velar-final roots in the database.