Turkish Electronic Living Lexicon (TELL)

Sharon Inkelas & Aylin Küntay & Ronald Sprouse & Orhan Orgun

Inkelas, Sharon & Küntay, Aylin & Sprouse, Ronald & Orgun, Orhan 2000. Turkish Electronic Living Lexicon (TELL). *Turkic Languages* 4, (###).

This paper introduces the Turkish Electronic Living Lexicon (TELL), a searchable lexical database of Turkish. The current version of TELL provides phonologically accurate transcriptions of the Turkish words that an actual native speaker recognized out of a larger master list of words culled from dictionaries and other print sources. TELL is accessible over the Internet via a search engine that permits users to search for potentially complex phonological patterns and to download and save their results. Designed primarily for academic research into Turkish, TELL also has obvious applications for students and teachers of Turkish.

Sharon Inkelas, Department of Linguistics, University of California, Berkeley, CA 94720, USA, E-mail: inkelas@socrates.berkeley.edu

Aylin Küntay, Department of Psychology, Koç University, Çayır cad., İstinye 80860, İstanbul, Turkey.

Orhan Orgun, Department of Linguistics, UC Davis, Davis, CA 95616, USA, E-mail: ocorgun@ucdavis.edu

Ronald Sprouse, Department of Linguistics, UC Berkeley, Berkeley, CA 94720, USA, E-mail: ronald@socrates.berkeley.edu

1. Introduction

This paper introduces the Turkish Electronic Living Lexicon (TELL), an ongoing project at the University of California at Berkeley which aims to establish a searchable lexical database of Turkish.[1] TELL is primarily designed for academic

[1] TELL was funded during 1995-1997 and is currently being funded through 2001 by US National Science Foundation awards #SBR-9514355 and #BCS-9911003 to

research into the phonological structure of Turkish but also has obvious applications for students and teachers of Turkish. TELL differs in content from standard print dictionaries of Turkish in providing phonologically accurate transcriptions of those Turkish words known to an actual native speaker. TELL is accessible over the Internet via a search engine that permits users to search for potentially complex phonological patterns and to download and save their results.

2. Motivation for TELL

The original motivation for TELL was to provide an accurate database for conducting phonological research into Turkish, which has long been an influential language in the development of phonological theory. Unfortunately, reliance on inadequate data has led a number of researchers into making dubious claims about Turkish. Unless a researcher has direct access to native speakers, which is not always the case, the researcher is forced to rely for hypothesis formation and testing on examples previously cited in the literature—potentially perpetuating errors—or on print dictionaries.

2.1. Inadequacies of print dictionaries as basis for phonological research

Many high-caliber dictionaries of Turkish exist. However, they are not only time-consuming to use but also inadequate for phonological research, for a number of reasons enumerated below.

2.1.1. Dictionaries are conservative

Dictionaries tend to be conservative, containing many words found in older literature but not known or used by the typical speaker. This problem is particularly acute for Turkish, due to the legacy of the highly artificial Ottoman literary language with its deliberate loans from Arabic and Persian. Many of these loans were restricted to elite literary style and probably never used in the everyday spoken language. They are, however, included in modern dictionaries. The native speaker represented in the

Sharon Inkelas. Funding for pilot studies was provided by the Abigail Hogden Publication Fund and the Hellman Faculty Research Fund, to whose funders the authors are grateful. TELL currently enjoys computational support provided by the University of California, Berkeley. Over the years the following individuals have worked for TELL: Jonathan Barnes, Andrew Dolbey, Gunnar Hansson, Dasha Kavitskaya, John B. Lowe, Yelda Mesbah. The authors also wish to thank Prof. Kemal Oflazer for valuable discussions and assistance.

TELL project, an educated man in his 60's, knew approximately half of the items in the second and third editions of the Oxford Turkish English dictionary. Words that speakers do not know are irrelevant to the computation of phonological generalizations about the synchronic form of the language. Yet the linguist perusing a dictionary cannot know which items to disregard on this basis.

2.1.2. Orthography is not sufficient

Dictionaries present words in orthography. Turkish orthography is close to phonemic, but does obscure the following four crucial phonological properties:

(1) Lateral and velar palatalization
Vowel length
Vowel epenthesis into initial clusters
Stress

The lateral /l/ and the velar plosives /k/ and /g/ all have palatal counterparts with which they contrast phonemically in the neighborhood of back vowels (all three are predictably palatal in the neighborhood of front vowels). Turkish orthography provides a means of indicating palatality: a circumflex on a vowel can indicate that the preceding consonant is palatal. However, the circumflex can also be used to indicate vowel length, making it ambiguous. This is illustrated by the following examples. ("Ox57" refers to the 2nd edition of the Oxford Turkish-English dictionary, published in 1957; "Ox92" refers to the 3rd edition, published in 1992. Here and elsewhere, pronunciations, presented in IPA, are those of the native speaker represented in version 1.0 of TELL.)

(2)	orthography	pronunciation	gloss	source
circumflex indicates velar palatalization:	*gâvur*	gʲɑvur	'infidel'	Ox57,Ox92
circumflex indicates vowel length:	*gâsıb*	gɑːsip	'usurper'	Ox57
circumflex indicates velar palatalization *and* vowel length:	*kâfi*	kʲɑːfi	'sufficient'	Ox57,Ox92

Not all forms in which a palatal consonant precedes a back vowel are spelled with a circumflex, however:

(3)	orthography	IPA	gloss	source
Contrastively palatal	*Hollanda*	holʲlʲɑndɑ	'Holland'	Ox92

consonant preceding back vowel, but no circumflex:	*meşgale*	meʃgʲɑle	'business'	Ox57, Ox92

In any case, the circumflex is falling out of use in contemporary written Turkish, so that even the words written with a circumflex in (2) are increasingly being rendered without one. In many cases, circumflexes present in the 2nd edition of the Oxford dictionary (e.g. *gâsıb* 'infidel') are omitted in the third edition (e.g. *gasıp*).

Even in conservatively spelled sources, there is no orthographic means at all of marking palatality on a consonant which is *not* followed by a vowel, as in these forms:

(4)	orthography	IPA	gloss	source
Contrastively palatal consonant following but not preceding back vowel; no circumflex	*vokal*	vokɑlʲ	'vocal'	Ox92
	makbul	mɑkbulʲ	'acceptable'	Ox57, Ox92

Some dictionaries, e.g. Ox57 and the Redhouse Turkish-English dictionary, indicate contrastive palatality on a word-final lateral or velar by listing the form that the accusative suffix takes for that word.

Also not represented well in the orthography is vowel length. As indicated above, the circumflex is used sporadically in the orthography to represent length. However, it is underutilized even in conservative spelling, as exemplified by the following words:[2]

(5)	orthography	IPA	gloss	source
Circumflex absent but vowel is long	kaza	kaza:	'accident'	
	tesir	te:sir	'impression, influence'	

Some dictionaries, e.g. Redhouse and Ox57, use nonorthographic symbols such as dashes in pronunciation guides to indicate vowel length. However, others (e.g. Moran's 1985 *Büyük Türkçe-İngilizce Sözlük*) simply fail to represent it at all. This

[2] The Ox57 dictionary represents some words with long vowels for which the TELL speaker (and other native speakers consulted) have short vowels. This "overrepresentation" of length presumably reflects conservative pronunciations or dialect variation.

is a great loss to the phonologist, since Turkish has a great many words (nearly 10%, according to recent TELL estimates) with phonemically long vowels.

A third area of pronunciation in which dictionary representations systematically differ from speakers' productions is in the rendition of words spelled with apparently tautosyllabic consonant clusters. Most speakers systematically break these up with epenthetic vowels, whose quality is of considerable interest to the phonologist. Yet dictionaries give the phonologist no indication that there is a vowel in these positions:

(6)

standard orthography	IPA	gloss	source
protesto	pɪrotesto	'protest'	Ox57
tren	tiren	'train'	Ox57, Ox92
streptokok	sitreptokok	'streptococcus'	Ox92
ansambl	ansambɪl	'ensemble'	Ox92

Finally, the orthography of Turkish does not mark stress. There are relatively few minimal pairs in Turkish which differ only in the position of stress. However, there are a number of items which follow neither of the regular stress placement rules (final stress, for ordinary words, and a more complex pattern of nonfinal stress, for place names and foreign names used in Turkish (see Sezer 1981, Inkelas 1999).

(7) Words with exceptional stress:

masa	[ˈmɑsɑ]	'table'
Bermuda	[berˈmudɑ]	'Bermuda'
tarhana	[tɑrˈhɑnɑ]	'dried curds'

These exceptional words play an important role in the stress system as a whole. Some dictionaries (e.g. Ox57, Redhouse) use nonorthographic symbols such as accent marks in pronunciation guides to indicate the position of stress, but others (e.g. Ox92, Moran) leave it out altogether.

2.1.3. Morphophonemics inadequately represented in many dictionaries

Another crucial aspect of the phonology of a Turkish word is the morphophonemic alternation pattern that it shows under suffixation. These fall into several types:

(8) *Morphophonemic properties of roots:*
Harmony pattern taken by suffixes
Vowel length alternations
Consonant length alternations
Consonant voicing

According to the rules of Vowel Harmony, harmonic suffixes appear with back vowels when the stem they attach to has a back vowel in its last syllable, and front vowels otherwise. High suffix vowels also agree in roundness with the closest stem vowel. Occasionally, however, stems violate this pattern by triggering disharmony on suffixes. This happens only with back vowel stems:[3]

(9) *Disharmony on suffixes (orthographic forms only)*

nominative	accusative	gloss	source
saat	*saat-i*	'hour'	Ox57, Ox92
dikkat	*dikkat-i*	'attention'	Ox57, Ox92
istimlak	*istimljk-i*	'expropriation'	Ox92
garb (garp)	*garb-i*	(the West; Europe'	Ox92 (ox57)

The Redhouse and both editions of the Oxford dictionary list the form that the accusative suffix (homophonous with the 3rd person possessive) takes for such words, as an indication that suffixes generally take front vowel harmony. Moran (1985) does not do this.

Some phonologically contrastive information within roots is neutralized in the citation form in which lexemes are typically listed in dictionaries. For example, since long vowels shorten in closed syllables, a word ending in an underlyingly long vowel followed by a consonant will shorten that vowel in citation form. Only the Redhouse dictionary marks such vowels as long.

(10) *Vowel length obscured in citation (nominative, for nouns) form*

nominative (orthography)	nominative (IPA)	accusative (IPA)	gloss
zaman	zɑmɑn	zɑmɑːn-ɨ	'time'

[3] While Clements & Sezer (1982:242) claim that some front-vowel roots might take back vowel suffixes, e.g. *fevk-ı* 'top-accusative' and *utarid-ı* 'Mercury-accusative', this phenomenon does not seem to stand up to additional scrutiny. We found, in a small study conducted with native speakers in Istanbul, that even the few speakers who exhibit back-disharmony in suffixed forms of these roots exhibit it *only* in accusative or possessed forms. For all other suffixes tested, e.g. the plural or non-accusative case endings, these same speakers exhibit front harmony in suffixes, i.e. *fevk-ler* (**fevk-lar*) 'top-pl' and *utarit-ten* (**utarid-dan*) 'Mercury-abl'. It is thus not a general morphophonemic property of these roots that they condition disharmonic suffixes (cf. the uniformly front-vowel conditioning roots such as *saat* 'hour', in (9) and (16)). The inescapable conclusion is that the accusative and possessed forms of *utarit* and *fevk* are simply suppletive.

nüfus	nyfus	nyfuːs-u	'people, souls'
mecnun	mediʒnun	mediʒnuːn-u	'madly in love'

For some reason, a parallel neutralization in the length of word-final consonants *is* marked systematically in dictionary pronunciation guides; the fact that the final consonant of *had* 'boundary' is actually a geminate is revealed by listing its accusative form (*haddi*) in the Oxford and Moran dictionaries.

2.1.4. Etymological info often not given

Since Lees (1961) the theoretical literature has seen many claims that the lexicon of Turkish is stratified, with different sectors of the vocabulary (typically native vs. nonnative) obeying different generalizations. This claim has been made as recently as Itô & Mester (1995). Unfortunately, however, most dictionaries do not provide the essential etymological information with which to test such claims.

2.1.5. Time-consuming to search

Even the most phonologically accurate print dictionary of a language as well-documented as Turkish poses the problem for the phonologist of providing so much information that a manual perusal of the whole dictionary to see how many forms of a particular phonological type occur is prohibitively time-consuming. The 2nd edition of the Oxford dictionary has over 16,000 entries; the 3rd edition, over 20,000. It is no surprise that studies of Turkish are not routinely accompanied by the kinds of dictionary counts seen for languages with much smaller dictionaries—or for languages like English or Spanish for which electronic dictionaries are readily accessible.

3. Desiderata for a lexical database

Given the difficulties enumerated above in using even the best print dictionary of Turkish to conduct phonological research, the first author proposed in 1995 to build an electronic dictionary of Turkish that included not only the contents of two excellent print dictionaries but also phonologically accurate transcriptions of the pronunciations of those forms by several native speakers. The primary desiderata for the database were as follows:

(1) The database should include a comprehensive—or at least a representative—list of words in actual use by speakers.
(2) The database should provide each word with a minimal morphological parse, to assist the nonnative speaker in isolating the root.
(3) The database should list the language of origin of the root in each word.

(4) The database should provide a phonological transcription of the word as pronounced by a native speaker.
(5) The database should provide morphophonemic information about each word, so that information about the underlying form can be recovered.
(6) The database should be searchable over the Internet, so that it can be used at no cost by linguists worldwide.

The next section describes TELL, the Turkish Electronic Living Lexicon, designed to meet these goals.

4. Structure of TELL

TELL was begun in 1996. The first version, TELL 1.0, was made public in 1998. TELL continues to be expanded and refined, and a second version is expected in a year or so. This paper describes version 1.0; novel components of version 2.0 are briefly sketched in a later section. TELL consists of four parts:

(1) Master list of dictionary headwords.
(2) Morphological roots of the headwords.
(3) Etymological information for the headwords.
(4) Phonological transcriptions of native speaker pronunciations of the headwords, in isolation and in combination with various suffixes.

These will be described in turn.

4.1. Master lexeme list

The master list of lexemes represented in TELL is a combination of three print sources: the 2nd and 3rd editions of the Oxford Turkish-English dictionaries, and place names from a PTT (Posta Telgraf Telefon) area code directory for Turkey and from a tour guide for Istanbul. Place name sources are not typically well-represented in dictionaries, yet are important to the phonologist primarily because of the distinctive stress pattern that they exhibit (see e.g. Sezer 1983). The PTT directory was comprehensive, but contained many place names not known to the native speaker from whom the words in the master list were to be elicited. The names from an Istanbul tour guide were included to increase the number of place names known to the native speakers whose knowledge is represented in TELL.

The two Oxford dictionaries were selected for four reasons. (1) Their lexical coverage is broad, with the 2nd edition containing more Arabic and Ottoman items and the 3rd edition more European loans. (2) Unlike many other dictionaries, they provide stress and some etymological information (2nd edition) and part of speech and semantic class (3rd edition). (3) They provide English translations, extremely

useful to the linguist who wishes to provide glosses for items extracted from TELL. (4) Both dictionaries have good print quality, making optical character recognition possible. (Competing dictionaries, e.g. the venerable Redhouse Turkish-English dictionary, had poor print quality as well as a fatal (for Optical Character Recognition purposes) mixture of Latin and Arabic characters.

With the kind permission of Oxford University Press, TELL was allowed to scan and perform optical character recognition of both dictionaries. The resulting texts were marked up using Standard Generalized Markup Language (SGML) language. SGML tags identify the elements in and logical structure of a text. For TELL's purposes, the following items were deemed relevant and tagged: Headword, pronunciation information (e.g. stress), part of speech, semantic class, gloss. An example entry from the 2nd edition, both before and after SGML markup, is shown below. <L> tags surround each lexeme; glosses are tagged with <G>. The headword, *ab*, is tagged as <HW>, while the subheadword, *~u hava* (interpreted as *abu hava*), is tagged as <X>:

(11) Entry in dictionary: ab Water; rain; river. ~u hava, climate.
SGML markup of entry: <ENTRY RN=″99960″ SRC=″OX57″>
<HW><L>ab</L> <G> (<STR>-</STR>) Water;
rain; river,</G></HW> <X> <L>~u hava</L>,
<G>climate.</G> </X> </ENTRY>

The number of headwords in the data is as follows:

(12)	2nd edition of the Oxford Turkish English dictionary	headwords: 17,000
	3rd edition of the Oxford Turkish English dictionary	headwords: 19,911
	Place names from Istanbul tour guide	headwords: 175
	PTT area code directory of Turkish cities	headwords: 4,728
	Total headwords:	41,834
	Total phonologically unique headwords (MASTER):	30,096

Once all the entries from the text sources were pooled and duplicates removed, the result was a list containing just over 30,000 lexemes. This list, termed MASTER, is the basis for the TELL database.

In order to make the data maximally accessible, the data are represented in an ASCII code which uses no platform-specific special characters. The table below indicates the correspondences:

(13)

orthography	TELL ASCII code	orthography	TELL ASCII code
a	a	l	l
â	a@	m	m
b	b	n	n
ç	c@	o	o
c	c	ö	o@
d	d	p	p
e	e	r	r
f	f	s	s
g	g	ş	s@
ğ	g@	t	t
h	h	u	u
ı	i@	ü	u@
i	i	û	u@@
î	i@@	v	v
j	j	y	y
k	k	z	z

Once MASTER was complete, the database was fleshed out in three orthogonal directions: morphological, etymological, and phonological.

4.2. Morphological root extraction

Many if not most of the words in MASTER are morphologically complex. There are two reasons for this. First, many entries in the Oxford dictionaries consist of a number of words all derived from the same root, with the alphabetically first derivative arbitrarily functioning as the headword. For example, the 2nd edition's entry for *gelincik* 'weasel' contains the subheadwords *gelinlik* 'quality of a bride' and *gelin havası* 'fine weather' (among others). All (including headword *gelincik* 'weasel, poppy') are derived from the root *gelin* 'bride'—yet it is the alphabetically first *gelincik*, not the other derivatives, which made it into MASTER. Second, many (if not most) place names in Turkish, of which TELL contains several thousand, are themselves morphologically complex, having literal meanings such as 'big black spring' (*Büyükkarapınar*) or 'with (an) oil lamp' (*Kandilli*).

For the phonologist using TELL, it is imperative to know whether a given word is a compound (as *Büyükkarapınar*) or contains suffixes (as *gelincik*), as many phonological phenomena are crucially conditioned by the morphological structure of the phonological string in question. Vowel harmony, for example, applies between stem and suffixes but does *not* apply between the two members of a compound. The phonologist searching for disharmonic vowel sequences in Turkish needs to know the morphological relationship between each pair of adjacent syllables; the

phonologist examining root structure constraints in Turkish needs to be able to isolate the root in each word.

Fortuitously for TELL, Prof. Kemal Oflazer of Bilkent University has developed a state-of-the-art morphological analyzer for Turkish, through which he was kind enough to run the then-current list of TELL words in 1996. Some 17,523 lexemes (nearly 60%) were recognized by the parser. The resulting roots exist in a list called ROOTS, which is linked to the MASTER list. Like the lexemes in MASTER, the roots in ROOTS are represented in standard orthography.

4.3. Etymologies from various dictionaries and articles

With the aim of equipping as many TELL entries as possible with etymological information, TELL researchers methodically went through a 5000-word etymological dictionary of Turkish (Eyuboğlu 1988) as well as numerous articles on the etymological origins of Turkish words (Ozon 1962, 1973; Püsküllüoğlu (1997; Stachowski 1975, Tzitzilis 1987). The languages claimed in these works to be the source of the lexemes in TELL were entered into a database called ETYMA, linked to MASTER. This methodology produced etymological identifications for 11,445 of the MASTER lexemes.

While the etymological dictionary was scoured in its entirety, this was not done with the articles, which were substantially more time-consuming to work through. Instead, TELL researchers concentrated their efforts on lexemes beginning with the following letters: [a, b, ç, c, e, f, i, ı, j, m, o, ö, p, t, u, ü, v]. The spread was intended to provide a reasonably representative sample of native vs. borrowed items.

Since the majority of sources consulted focused on loans in Turkish, the set of etymologically identified items is therefore heavily tilted toward borrowings. Nonetheless, it provides a more comprehensive etymological picture of Turkish than any of the comprehensive print dictionaries.

4.4. Pronunciations from one native speaker in various morphological contexts so that morphophonemic properties are revealed

The most novel feature of TELL, and the feature most important to the phonologist using the database, is the inclusion of pronunciation information for each orthographically represented lexeme. During the summers of 1996, 1997 and 1998, elicitation from a native speaker was conducted in Istanbul. The first speaker selected for the TELL project was a 63-year old college-educated male who had lived in Istanbul his entire life.

The speaker was presented with a randomized list of all of the lexemes in MASTER, minus those suffixes, acronyms and abbreviations that it was possible to weed out in advance. The speaker was asked to pronounce only those items which he

knew and used. (TELL was not interested in "reading pronunciations" of unfamiliar words.) Moreover, the speaker was asked to pronounce each lexical item not only in its isolation form but also in several different morphological contexts. This was done in order to reveal any morphophonemic alternations in the root.

Nominals were elicited in the nominative (= dictionary citation) form, as well as in the accusative, "professional", 1st person singular possessive and 1st person singular predicative. Verbs were elicited in the long infinitive (= dictionary citation) form, as well as in the aorist and in the causative.[4]

(14) Examples of elicitation: nominals

citation form (orthographic)	gloss	transcribed pronunciations (IPA)				
nominals:		nom.	acc.	prof.	1sg poss.	1sg pred.
yol	'way'	jol	joˈlu	jolˈʤu	joˈlum	ˈjolum
araba	'car'	araba	arabaˈjɨ	arabaˈʤɨ	araˈbam	ArAˈbAjµm

Examples of elicitation: verbs

citation form (orthographic)	gloss	Transcribed pronunciations (IPA)			
verbs:		citation	long infinitive	aorist	causative
etmek	'do'	etˈmek	etˈmek	eˈder	etˈtir
aktarma	'transfer'	aktarˈma	aktarˈmak	aktaˈrɨr	aktarˈtɨr

The five morphological contexts for nouns and three (or four) for verbs were selected on the basis of a pilot study using native speakers in Berkeley. The vowel-initial allomorphs of the accusative, 1st singular possessive, 1st singular predicative and aorist suffixes reveal underlying properties of stem-final consonants which may

[4] The 2nd edition of the Oxford dictionary sometimes cites verbs in the long infinitive (e.g. *çakışmak*, 'fit into one another'), and sometimes in the short infinitive (e.g. *ağlama* 'murmuring of water'). In the latter case, the speaker pronounced both the short infinitive (= dictionary citation) as well as the long infinitive, aorist and causative.

The speaker represented in version 1.0 of TELL sometimes produced causative stems in the imperative, as *et-tir* 'do-causative', but more often in the aorist, as *aktar-t-ır* 'transfer-causative-aorist' (12). To stems whose dictionary citation form already contained a causative suffix, the speaker supplied a second causative, as *hızlandırmak* 'to accelerate' → *hızlan-dır-t-ır* 'accelerate-causative-causative-aorist'. These were invariably produced in the aorist.

otherwise be neutralized in the citation form of the stem. This is true, for example, of the root *ecdat* 'ancestors', whose accusative form reveals underlying vowel length and final consonant voicing. For verbs, the aorist context was employed to uncover the underlying properties of root-final consonants. The root *et-* 'do', for example, displays final consonant voicing before the aorist suffix:

(15)

citation	accusative / aorist	
ecdat [edʒdɑt]	*ecdadı* [edʒdɑːd̪ɨ]	'ancestors(-acc)'
etmek [etmek]	*eder* [eder]	'do(-inf/-aorist)'

The 1st singular possessive was included to provide more information on roots triggering disharmony on suffixes. Such roots, e.g. *saat* 'hour', have back vowels in their final syllable yet trigger front harmony in suffixes, e.g. the accusative (*saat-i*) (Clements & Sezer 1982). In the literature it is assumed that these roots trigger front harmony in *all* harmonic suffixes, not just the accusative (and/or homophonous 3rd possessive), which is most commonly cited. This is certainly true for *saat*, as the following suffixed forms exemplify:

(16) Behavior of *saat* 'hour', standard for all speakers:

citation	*saat*	sɑ.ɑt
accusative / 3sg. possessive	*saati*	sɑ.ɑ.ti
1sg. possessive	*saatim*	sɑ.ɑ.tim
plural	*saatler*	sɑ.ɑt.ler
professional	*saatçi*	sɑ.ɑt.tʃi
abstract noun	*saatlik*	sɑ.ɑt.lik

However, pilot studies conducted by TELL show that other, less frequently used roots cause disharmony *only* on the accusative / 3sg possessive suffixes:

(17) Behavior of some speakers in pilot study conducted by TELL:

Citation	*işt̪irak*	iʃt̪irɑk
Accusative	*işt̪iraki*	iʃt̪irɑːkɨ
1sg.possessive	*işt̪irakɨm*	iʃt̪irɑːkɨm

Since this phenomenon had not been previously reported in the literature, the opportunity was taken to see how pervasive (if at all) it is.

The 1st singular predicative was included because of its distinctive pre-stressing pattern. The "professional" suffix was included because it is uniformly consonant-

initial. For verbs, the causative was included because it (along with the aorist) shows considerable allomorphy and is interesting in its own right.

The native speaker consultant was familiar with 17,593 of the 30,096 items in MASTER. Of these, 1934 are verbs and 15,591 are nominals. Taking into account the various morphological forms that were elicited, the speaker pronounced some 85,000 forms. The pronunciations were recorded on analog audiotape on an inexpensive Walkman-style tape recorder and transcribed by a native speaker. The transcriptions, which were phonemic, were rendered in ASCII phonemic transcription system capable of expressing all phonologically contrastive features of Turkish. The transcription system is presented below:[5]

(1) ASCII transliteration of phonemic transcriptions:

TELL ASCII code	IPA	TELL ASCII code	IPA
a	A	l	l
b	b	l@	lJ
c	dʒ	m	m
c@	tS	n	n
d	d	o	o
e	e	o@	ø
f	f	p	p
g	g	r	r
g@	g^j	s	s
h	h	s@	ʃ
i@	ɨ	t	t
i	i	u	u
j	ʒ	u@	ü
k	k	v	v

[5] Velar and lateral palatality are transcribed only when phonetically unexpected. Velars are predictably palatal in Turkish when tautosyllabic with a front vowel follows (e.g. *kek* 'fruitcake' [k^jek^j]); laterals are predictably palatal when adjacent to a front vowel (*lig* 'league' [l^jig], *fil* 'elephant' [fil^j], *bela* 'trouble' [bel^{j}ɑ:]). TELL does not transcribe this redundant palatality, reserving the palatal symbol for phonetically unconditioned palatality (e.g. *gavur* 'infidel' [g^javur], transcribed in TELL as "g@avur"). Lateral palatality is also predictable word-initially when /ɑ/ follows, e.g. *lale* 'tulip' [l^jɑ:l^je]. However, since this palatalization is phonetically unusual, and since speakers consulted in pilot studies for the TELL project had exceptions to the generalization (e.g. *lala* [lɑlɑ] 'servant', TELL marks palatality on word-initial laterals (thus transcribing *lale* as "l@a@le").

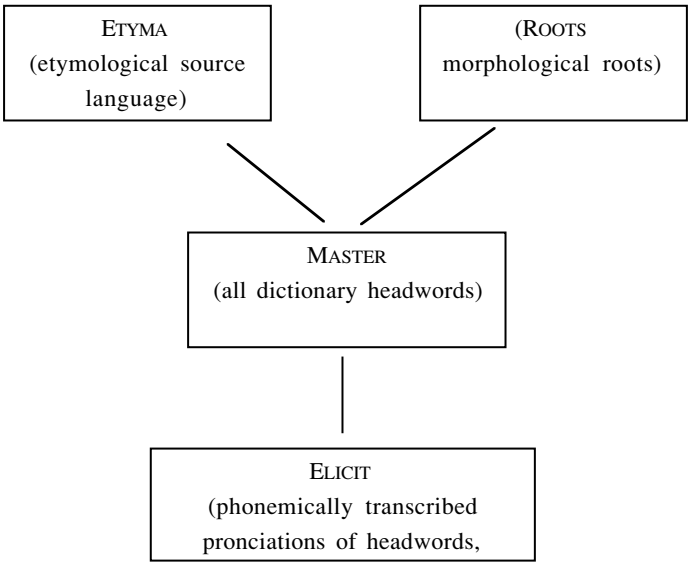
k@	k^j	y	j
		z	z

Stress was also transcribed for all forms elicited from the native speaker. The TELL convention for marking stress is to use a single quote following the stressed vowel. Thus final-stressed *kitap* 'book' is transcribed as "kita'p", and initial-stressed *masa* 'table' as "ma'sa". Vowel length is transcribed with a colon following the vowel; thus *kaza* 'accident' [kɑzɑ:] is transcribed in TELL as "kaza:'".

The approximately 85,000 transcriptions exist in a database termed ELICIT, which is linked to MASTER.

User interface of TELL

The table below summarizes the structure of TELL:



The four datatypes exist as four Microsoft Access™ tables, linked by a common index. This structure permits the user to search, for example, for all words that simultaneously meet a given orthographic description, have a certain pronunciation, derive from a particular source language, and for which a root meeting a given description has been extracted. However, Microsoft Access™ is insufficient to perform the more sophisticated searches that a phonologist might require. In particular it does not support regular expressions, crucial to the definition of natural classes (e.g. front vs. back vowels, or voiced vs. voiceless consonants, or heavy vs. light syllables). Therefore, a special search engine was designed for TELL that would permit users to access TELL over the Internet and search its contents in multiple ways. Written in Perl, the engine lives on a Unix server and operates on a version of the TELL data stored in Berkeley Database format. The search engine is accessed via the TELL web site and has a web interface permitting the user to configure each individual search.[6]

[6] http://socrates.berkeley.edu:7037/TELLhome.html

Search parameters in TELL

The TELL search engine provides the user with a number of options in defining a search:

(2) Search parameters of the TELL search engine

(19) Data to search
- Text sources:
  - *Dictionary headwords only*
  - *Place names only*
  - *Both dictionary headwords and place names*
- Etymological restrictions:
  - *search all words, or restrict search to native words, to nonnative words, or to words originating from a particular language*
- Morphological restrictions:
  - *search all words, or only those for which roots have been extracted*
- Representations to search:
  - *orthographic (dictionary/place name entries), or phonemic (elicited)*
- Fields to search:
  - *accusative, 1sg.possessive, professional, 1sg.predicate, infinitive, aorist, or causative (phonemic representations only), plus citation (both orthographic and phonemic representations)*
- Expression to search for:
  - *any regular expression*

More than one field can be searched simultaneously; in such cases, specifications are conjunctive. For example, the following search

(20) Field 1: [k>] [citation]
Field 2: [ei>] [accusative]

will return all nominals whose pronunciations end in [k] (the ">" means "word boundary") in the nominative (=citation) and end in the sequence [ei] in the accusative.[7] These would include stems like *bebek* 'baby', whose final velar drops out in before a vowel-initial suffix, as in the accusative *bebeğ*i [bebei].

[7] Perl's built-in word-boundary metacharacter '\b' does not produce the correct results, as it erroneously matches the boundary between regular letters and the diacritic

Another possible search might combine orthographic and phonemic descriptions. The Lexeme field contains the orthographic representation of the citation form. Thus, the following search

(21) Field 1: [>pr] [lexeme]
Field 2: [>pi@r] [citation]

will return all words which are spelled with an initial "pr" cluster which, in pronunciation, is broken up by an epenthetic high back vowel.

Since the TELL search engine supports regular expressions, the user can transcend these pedestrian searches and seek broader patterns in the data. For example, the regular expressions in the following search:

(22) Field 1: [kg] [citation]

return all words ending in a velar consonant ([k] or [g]); the following search

(23) Field 1: [kg]@[i@|(u[^@])|(e)|o[^@] [citation]

returns all forms with palatal velars which precede back vowels.

Because regular expressions can be onerous to compute and type, TELL has a "metacharacter" utility that allows users to use predefined characters to stand for a fixed set of regular expressions.[8] These are currently built in to the search engine:

metacharacter	regular expression	phonological characterization
C	(?:[bcdfghjklmnprstvyz]@?)	# CONSONANTS
S	(?:[hlmnry]@?)	# SONORANT CONSONANTS
O	(?:[bcdfgjkpstvz]@?)	# OBSTRUENTS
G	(?:[bcd]|g@?)	# VOICED STOPS AND AFFRICATES
K	(?:c@|[ptk]@?)	# VOICELESS STOPS AND AFFRICATES
V	(?:[aeoui]@?)	# VOWELS
I	(?:[ui]@?)	# HIGH VOWELS
R	(?:[ou]@?)	# ROUND VOWELS

symbol '@', which is normally nonalphabetic. As a result, we defined '>' to treat '@' as alphabetic.

[8] These are separate from Perl's built-in metacharacters and in some cases supplant them, e.g. '>' replaces '\b' as the word-boundary metacharacter.

E	(?:[ie]|[ou]@)	# FRONT VOWELS
A	(?:[uao]|i@)	# BACK VOWELS
B	(?:[pvbfm])	# LABIAL CONSONANTS
T	(?:[cdghjklnrstyz]@?)	# NON-LABIAL CONSONANTS
U	(?:u[^@]|u@@)	# for orthographic lexeme field in MASTER
>	(?:^| |tl$)	# word boundary

Thus, the search conducted in (23) can be triggered by the following:

(24) Field 1: [kg]@A | citation

which is much simpler to type and far less prone to error. Advanced users can define new metacharacters as needed to further simplify their search expressions.

Vowel length and stress can be searched for by invoking the colon and single quote that mark these features in the TELL transcription. Thus

(25) Field 1: >C*V'.*V: | citation

returns all forms with initial stress and a noninitial long vowel.

Display and saving of search results

TELL automatically displays search results in the form of a table. The example in (26) illustrates the results of a search for all citation forms containing a sequence matching the regular expression "eC*VC*u@", meaning all forms containing the vowel [e], then some string of consonants including at least one labial, then the vowel [ü]. Displayed, at user request, are the citation, lexeme, etymology and accusative fields:[9]

(26)

citation	lexeme	etymology	accusative
ecis@bu@cu@s@	ecis@ bu@cu@s@		ecis@bu@cu@s@u@
c@es@mibu@lbu@l	c@es@mibu@lbu@l		c@es@mibu@lbu@lu@
tribu@n	tribu@n		tribu@nu@
entipu@ften	entipu@ften		entipu@fteni
ilmu@haber	ilmu@haber	Ar	ilmu@haberi
manipu@lato@r	manipu@lator	Fr	manipu@lato@ru@
difu@:ze	difu@ze		difu@:zeyi
simu@ltane	simu@ltane		simu@ltaneyi

[9] "Ar" = Arabic, "Fr" = French, "Yun" = Greek

okaliptu@s	okaliptu@s	Yun	okaliptu@su@
tifu@s	tifu@s	Yun	tifu@su@
dinibu@tu@n	dinibu@tu@n		dinibu@tu@nu@

Advanced users also have (by permission of TELL) the option of viewing a tab-separated text file containing the search results; from the latter, it is easy to download results to the user's home computer. Advanced users also have the option of saving search results on the TELL server. The advantage of this is that the saved results of prior searches can then be searched again.

Because of space limitations on the TELL server at the time of this writing, non-advanced users are limited to seeing the first 100 items in any set of search results, although it is hoped that this limit will be raised in the future. The user is told how many matches were found, even when not all can be displayed.

The user has a variety of options in determining how search results are displayed. Any of the fields in which search expressions can be typed—root, lexeme, citation, accusative, etc.—are available as display options as well. Thus, if the user is searching for all words ending with a velar in the citation form, the logical default would be to display the citation forms in the search results. However, it is equally possible to display only the accusative forms of words meeting the description of the search expression—or, for that matter, to display those words in all of their forms. The user also has the option of displaying the morphological root and etymological source language (if available) of all words found by the search expression (see example (26)).

Results of TELL

Though most of the TELL research team's efforts have thus far gone into building the database and search engine, a number of findings have already been made. For example, TELL has permitted testing of the following two claims made in the literature about Turkish:

> *Schein & Steriade (1986: 714)*: Turkish lacks monomorphemic geminates. *TELL*: Turkish has over six hundred roots containing geminate consonants
>
> *v. d. Hulst and v. d. Weijer (1991: 13)*: vowel length in Turkish is marginal. *TELL*: almost 3,000 words, or 16% of the elicited forms in TELL, contain phonemically long vowels. (Only a small fraction of these long vowels correspond, in orthography, to a short vowel-soft g sequence, as in *dağ* [dɑː] 'mountain').

In a joint paper by members of the TELL team, Inkelas, Hansson, Küntay & Orgun (1998) used TELL to test the empirical validity of the claim made by Lees (1966)

and defended by Foster (1969) that Turkish subscribes to a constraint of labial attraction. Labial attraction supposedly rounds high back vowels when separated from an /ɑ/ in the preceding syllable by some number of consonants that includes at least one labial. Labial attraction is in competition with vowel harmony, which predicts an /ɨ/ in that same position. Inkelas et al. were able to show, using TELL data, that Labial Attraction is not a statistically valid generalization over Turkish. This confirms the conclusions of Zimmer (1969) and Clements & Sezer (1982) that Labial Attraction is too exception-ridden to be a true rule of Turkish. This paper also uses TELL's etymological feature to challenge the narrower claim of Ní Chiosáin & Padgett (1993), Itô, Mester & Padgett (1993), and Itô & Mester (1995) that Labial Attraction holds only within the native vocabulary of Turkish. A search of TELL revealed that Labial Attraction is actually stronger within *nonnative* items, presumably due to the fact that most the languages from which Turkish has borrowed most heavily contain the vowels /a/ and /u/ but not the vowel /ɨ/.

Work on the empirically elusive phenomenon of emphatic reduplication has been furthered by the TELL database. Yu (1998) and Wedel (forthcoming) used TELL to increase substantially the size of the corpus of emphatic reduplicated adjectives (e.g. *ter-temiz*, 'very clean'), on the basis of which they formulated new generalizations about this word-formation process.

Inkelas (forthcoming) uses TELL to examine intervocalic velar deletion. TELL shows that there are a number of exceptions to this well-known and highly productive process (e.g. *demagog* [demɑgog] 'demagogue', *demagog-u* [demɑgogu] 'demagog-acc', rather than the expected **demagog(u* [demɑgo.u]). Furthermore, the exception rate varies by morphological category, with the predicative suffix more likely to preserve a preceding intervocalic velar than the possessive.

Future of TELL

TELL is presently in its second phase of funding, and has goals that go far beyond the goals of the first phase. While the work of the first phase will be continued—adding more speakers, finishing the root extraction and etymological research, providing English translations and part of speech information for existing lexemes—Phase 2 of the TELL project has the following new aims:

(1) Link TELL to text corpus
(2) Link TELL transcriptions to audio files

By linking the TELL database to an electronic text corpus of Turkish, TELL will be enhanced in the following ways. First, text frequency of each lexeme in the database can be estimated. Text frequency has recently been shown to useful in estimating

morphological productivity (Baayen 1993) and psychological salience of phonological patterns (Frisch & Zawaydeh forthcoming). Second, the syntactic and semantic contexts in which items appear can be evaluated and concordances can be provided. This will be of use not only to the syntactician and semanticist but also to the language learner.

In Turkish, of course, due to the highly suffixing nature of its morphology, root frequency may be of equal or greater interest than word frequency. The linguist interested in the distribution of disharmonic roots is interested in how many times a speaker is likely to be exposed to words containing the disharmonic root *anne*, rather than how many times a speaker is likely to be exposed to any particular derived or inflected form of that root. TELL will thus tabulate both word *and* root frequency for Turkish.

The second main goal of the second phase of the project is to provide audio files for each transcription in the TELL database. This will be done not for the speaker currently represented in TELL, whose audio recordings are not of sufficiently high quality, but rather for the second and third speakers whose data is currently being processed and will soon be added to the database. These speakers were recorded on digital tape in soundproofed rooms. Users of TELL will be able to listen to (or download) high-quality recordings of the words that their searches return. This utility will serve phonologists who wish to check TELL transcriptions, phoneticians who wish to study particular sounds of Turkish, and language learners who wish to hear the words they are learning pronounced by a native speaker.

References

Alkım, V. Bahadır et al. (eds.) 1968. *Redhouse yeni Türkçe-İnglilizce sözlük*. 13th edition. İstanbul: Redhouse Yayınevi.

Baayen, Harald 1993. On frequency, transparency and productivity. In G. Booij & J. van Marle (eds.) *Yearbook of morphology* 1992. 181-208.

Clements, G. N. & Sezer, E. 1982. Vowel and consonant disharmony in Turkish. In: van der Hulst, H & Smith, N. (eds.) *The structure of phonological representations* 2. Dordrecht: Foris. 213-255.

Eyüboğlu, I. Z. 1988. *Türk dilinin etimoloji sözlüğü*. İstanbul: Sosyal Yayınlar.

Foster, J. F. 1969. *On some phonological rules of Turkish*. [PhD dissertation, University of Illinois at Urbana-Champaign.]

Frisch, Stefan & Zawaydeh, Bushra (forthcoming). The psychological reality of OCP-Place in Arabic. To appear in *Language*.

Hony, H. C. & İz, Fahir 1957[2]. *A Turkish-English dictionary*. Oxford: Clarendon Press.

Inkelas, Sharon 1999. Exceptional stress-attracting suffixes in Turkish: representations vs. the grammar. In: Kager, R. van der Hulst, H. and Zonneveld, W. (eds.) *The prosody-morphology interface*. Cambridge: Cambridge University Press. 134-187.

Inkelas, Sharon & Hansson, Gunnar Küntay, Aylin & Orgun, Orhan 1998. Labial attraction in Turkish: an empirical perspective. Paper presented at the 16th International Conference on Turkish Linguistics, Oxford University.

Itô, Junko & Mester, Armin 1995. Japanese phonology. In: Goldsmith, John (ed.) *The handbook of phonological theory*. Blackwell. 817-838.

Itô, Junko & Mester Armin, & Padgett, Jaye 1993. Licensing and redundancy: Underspecification in Optimality Theory. Linguistics Research Center report #LRC-93-07, University of California, Santa Cruz.

İz, Fahir, & Hony, H. C. & Alderson, A. D. 1992[3]. *The Oxford Turkish-English dictionary*. Oxford: Oxford University Press.

Lees, Robert 1961. *The phonology of Modern Standard Turkish*. Bloomington: Indiana University Publications.

Lees, R. B. 1966. On the interpretation of a Turkish vowel alternation. *Anthropological Linguistics* 8, 32-39.

Moran, A. Vahid 1985. *Büyük Türkçe-İngilizce sözlük*. Istanbul: Adam Yayınları.

Ní Chiosáin, Máire & Padgett, Jaye 1993. Inherent Vplace. Linguistics Research Center report #LRC-93-09, University of California, Santa Cruz.

Özön, M. N. 1962. *Türkçe yabancı kelimeler sözlüğü*. İstanbul: İnkılap ve Aka Kitabevleri: Tan Gazetesi ve Matbaası.

Özön, M. N. 1973[5]. *Büyük Osmanlıca Türkçe sözlük*. İstanbul: İnkılap ve Aka Kitabevleri.

Püsküllüoğlu, A. 1997. *Türkçedeki yabancı sözcükler sözlüğü*. Ankara: Arkadaş Yayınevi.

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

Sezer, Engin 1981. On non-final stress in Turkish. *Journal of Turkish Studies* 5, 61-69.

Stachowski, Stanislaw 1975. *Studien über die Arabischen Lehnwörter in Osmanisch-Türkischen*. Wroclaw: Zaklad Narodowy im. Ossolinskich.

Tzitziles, Christos 1987. *Griechische Lehnwörter im Türkischen: Mit besonderer Berücksichtigung der anatolischen Dialekte*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.

van der Hulst, Harry & van de Weijer, J. 1991. Topics in Turkish phonology. In Boeschoten, H. & Verhoeven, L. (eds.) *Turkish linguistics today*. Leiden: E. J. Brill. 11-59.

Wedel, Andrew (forthcoming). Perceptual distinctiveness in Turkish emphatic reduplication. To appear in the proceedings of the 19th West Coast Conference on Formal Linguistics.

Yu, Alan 1999. Dissimilation and allomorphy: the case of Turkish emphatic reduplication. [Poster presented at the 29th North East Linguistics Society meeting.]

Zimmer, Karl 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language* 45, 309-321.