Turkish Electronic Living Lexicon (TELL): A Lexical Database

Sharon Inkelas*, Aylin Küntay[†], C. Orhan Orgun[‡], Ronald Sprouse*

* University of California at Berkeley Department of Linguistics Berkeley, CA 94720 USA inkelas@socrates.berkeley.edu Ronald@uclink.berkeley.edu

> [†]Koç University Department of Psychology Istinye, Istanbul 80860 Turkey akuntay@ku.edu.tr

[‡] University of California at Davis Department of Linguistics Davis, CA 95616-8177 ocorgun@ucdavis.edu

Abstract

The purpose of the TELL project is to create a database of Turkish lexical items which reflects actual speaker knowledge, rather than the normative and phonologically incomplete dictionary representations on which most of the existing phonological literature on Turkish is based. The database, accessible over the internet, should greatly enhance phonological, morphological, and lexical research on the language.

The current version of TELL consists of the following components:

- Some 15,000 headwords from the 2d and 3d editions of the Oxford Turkish-English dictionary, orthographically represented.
- Proper names, including 175 place names from a guide of Istanbul, and 5,000 place names from a telephone area code directory of Turkey.
- Phonemic transcriptions of the pronunciations of the same headwords and place names embedded in various morphological contexts. (Eliciting suffixed forms along with stems exposes any morphophonemic alternations that the headwords in question are subject to.)
- Etymological information, garnered from a variety of etymological sources.
- Roots for a number of morphologically complex headwords.

The paper describes the construction of the current structure of the TELL database, points out potential questions that could be addressed by putting the database into use, and specifies goals for the next phase of the project.

1. Introduction

The goal of this paper is to introduce an ongoing project at the University of California at Berkeley aiming to establish a searchable lexical database of Turkish, usable by researchers at large as a resource for linguistic research and second language learners for pedagogical reference. The database, which we dub TELL (Turkish Electronic Living Lexicon), represents both dictionary information and actual speaker pronunciations of two native speakers. Turkish has long played a central role in phonological and morphological theories because of its interesting properties such as vowel harmony and agglutinative morphology, but the data used for previous analyses of the language have never been comprehensive. In addition, serious shortcomings of some of the research practices such as overreliance on archaic forms, and disregard for exceptional forms and idiolectal differences have led to dubious generalizations based on faulty data. Thus, there is a need for a comprehensive lexical corpus of Turkish that will permit a reliable testing of the existing claims about the lexical structure of the language and lead to new hypotheses that can be verified or falsified statistically. By providing access to different researchers of different theoretical bends, a common database against which predictions could be evaluated avoids the problem of pooling in disparate and often incompatible data sources for comparative purposes. An electronic lexical corpus provides a representative list of forms, unbiased with respect to the particular question at hand, on which frequency measures and statistical tests could be performed.

The first phase of the TELL project was started in 1995, supported by a grant from the US National Science Foundation (SBR-95-14355) to Sharon Inkelas. TELL had the following goals:

• To put together a master list of Turkish lexical items by computerizing two print dictionaries,

and place names from a telephone directory and a guidebook

- To elicit the items in the master list in different morphological contexts from native speakers and provide a phonemic transcriptions of the pronunciations
- To provide etymological information and morphological annotation for each of the lexical items in the database
- To build a web-based interface for TELL to allow both segmental searches within words and look-ups of glosses and etymologies.

The next section will present the motivations for developing an electronic database of Turkish lexical items that is augmented with phonological, etymological, and morphological annotations. Then the steps that led to the current structure of the tool will be described. Finally, goals for future development of the TELL project will be briefly characterized.

2. Motivations for different properties of TELL

TELL, in addressing the need for a widely accessible and accurate source of data on the lexical phonology of Turkish, has taken up in providing phonological, etymological, and morphological information for the lexical items acquired from print sources. The following three sections discuss the need for these expansions.

2.1. Need for a phonological database

Although the Turkish orthographic system can be said to be phonemic for general purpose use, it does not does not encode all relevant phonemic distinctions that are crucial for linguistic analysis. Standard dictionary pronunciation guides are also not up to task in marking features such as lengthening, palatalization, and epenthesis of vowels into onset clusters. The following table displays some of the phonology-orthography discrepancies we observe in our data:

Standard orthography	Actual pronunciation by TELL speaker	Gloss				
Epenthesis into onset clusters						
protesto	pirotesto	'protest'				
ansambl	ansambıl	'ensemble'				
	Ambiguous vowel length symbols					
gâvur	g ^j avur	'infidel'				
kâfi	k ^j a:fi	'sufficient'				
	Overrepresentation of vowel length	1				
râkib	rakib	'mounted'				
lâla	lala	'servant in charge of boy'				
	Underrepresentation of vowel lengt	h				
kaza	kaza:	'accident'				
tesir	te:sir	'impression, influence'				
Underr	epresentation of velar and lateral pal	atalization				
Hollanda	hol ^j l ^j anda	'Holland'				
meşgale	meşg ^j ale	'business'				

Table 1: Examples of discrepancies of orthography and pronunciation

By filtering all the headwords originating in speaker print sources through native pronunciations, TELL ensures phonological the accuracy of lexical representations. Consequently, linguists using the TELL database will not need access to native speakers for doublechecking the actual pronunciation of lexical items if they conduct their searches within phonetically transcribed lexemes.

2.2. Need for etymological annotation of the database

Another shortcoming of standard dictionaries is that they do not typically identify loans or list source languages, although etymological information is crucial for the analyst of Turkish. Many analyses of Turkish grammar, originating with Lees in 1966 and as recently as Ito and Mester (1995), have claimed that native and loan vocabulary are governed by different phonological systems. However, information about which items are indeed native is hard to come by, resulting in misstatements in the literature. (For example, Ito & Mester (1995) and Ní Chosáin & Padgett (1993) claim that Labial Attraction, a constraint claimed by Lees (1966) to require high vowels to be round if preceded by a labial consonant and the vowel [a], holds only of native roots. But of the seven supposedly native items they cite to illustrate the application of the constraint, TELL shows that at least three are loans).

The TELL project aims to equip as many TELL entries as possible with etymological information to enable the researcher to discover true generalizations about the differences (if any) between native and nonnative grammatical components in the language.

2.3. Need for morphological annotation of the database

Turkish-English dictionaries are organized conceptually around a root, and consist of alphabetically ordered subheadwords that are derivatives of that root. Rendering a morphological parse of the morphologically complex lexemes is critical to the phonologist who wishes to investigate root structure conditions or to the morphologist who wishes to find all derived words containing а certain root. Insofar as morphophonemic alternations are sensitive to morpheme boundaries, morpheme boundaries should be encoded in the database.

To extract the roots of complex lexemes, TELL project has employed a state-of-the-art morphphological analyzer developed by Prof. Kemal Oflazer of Bilkent University.

3. The Current Structure of TELL

This section will review the steps taken to construct the TELL database and the web-based search interface.

3.1. Creation of the master list

To form the Master List, headwords were extracted from the following sources through scanning, optical character recognition, and SGML-tagging. (The TELL project is grateful to Oxford University Press for permission to scan the Oxford Turkish-English dictionaries.)

- Oxford Turkish-English Dictionary, 2d edition (1957): 17,002 headwords
- Oxford Turkish-English Dictionary, 3d edition (1992): 19,911 headwords
- PTT area code directory for Turkey (providing placenames): 4,728 headwords
- Guidebook for Istanbul (providing placenames): 175 headwords

The total of phonologically unique headwords in the Master List is 30,096.

The internal structure of the dictionary entries were parsed out by SGML tagging that indicates headword, stress, part of speech, semantic class, and gloss. An example entry 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>:

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 headwords from the SGML-tagged entries were later extracted into a database program that allowed clean-up, removal of the duplicates, and transportation to the field for elicitation of native speaker pronunciations.

3.2. Elicitation and transcription

The 30,096 headwords in the Master list were elicited from a 63-year old male native speaker of Standard Istanbul Turkish and transcribed. The speaker knew 17,593 of the items, which were phonemically transcribed and marked for stress and consonant palatalization where unpredictable. As the following table indicates, certain phonological contrasts are neutralized in citation forms of roots:

Contrast	Citation	Accusative	Gloss
Vowel length before	zaman	zama:n-1	'time'
final consonant	saman	samanı	'hay'
Final consonant	kanat	kanad-1	'wing'
voicing	sanat	sanat-1	'art'
Suffix disharmony	saat	saat-i	'hour, watch'

Table 2: Phonological contrasts which emerge under suffixation

In order to detect such neutralizations of contrasts and reveal root phonology, nominals and verbs were elicited in several different morphological contexts. Nominals were elicited in the nominative (=citation) form, as well as in the accusative, first person singular possessive, first person singular predicative, and "professional" forms. Verbs were elicited in the long infinitive (=citation) form, as well as in the aorist and in the causative infinitive.

The elicited pronunciations were transcribed using an ASCII phonemic transcription system capable of rendering all phonologically contrastive features of Turkish including the properties that are not orthographically represented such as stress, lateral and velar palatalization. The transcriptions were subsequently converted into a database that was linked up to the Master List.

3.3. Identification of Etymological Source

Using a variety of print sources (Eyüboğlu, 1988; Ozon, 1962, 1975; Püsküllüoğlu, 1997; Stachowski, 1975; Tzitzilis, 1987), TELL researchers traced the etymological source languages for 11,445 headwords in the Master List. 15 different languages were recorded as source languages from which lexemes have been borrowed into Turkish. 2201, or 19% of the identified words are native, whereas the bulk (81% of the identified words) are nonnative. (This ratio of course reflects the skewing of the sources, most of which were works on loanwords in Turkish.) For borrowed words, the language(s) of origin were noted as a separate field in the database.

3.4. Morphological parsing

In order to extract roots from morphologically complex headwords, the morphological analyzer developed by Prof. Kemal Oflazer at Bilkent University was run on the TELL headwords. The roots of around 60% of the headwords (17,523) that were recognized by the parser were extracted and linked up to the Master List. Many of the unanalyzable entries are compounds, which the morphological analyzer treated either as as monomorphemic roots or (in some cases) did not recognize. These and other items with unrecognized roots remain to be dealt with in future passes on the data.

3.5. Constructing the search engine

The above processes yielded four databases (MASTER, ELICIT, ETYMA, AND ROOT). Since a user might easily want to search multiple database simultaneously, the four databases were linked together with a common index. Given this linking, a user can, for example, obtain access to all and only the elicited native items with disharmonic roots, by simultaneously searching ELICIT (for the elicited pronunciations), ETYMA (for the native distinction) and ROOT (for restricting the search to roots).

Permitting sophisticated searching of the multidimensional database required development of a specialized search engine. Standard commercial database management programs (such as Microsoft AccessTM) are inadequate to serve the needs of phonologists who wish to conduct sophisticated segmental or prosodic searches. In

particular, standard databases do not support regular expressions, which are crucial to the definition of natural classes (e.g., front vs. back vowels, or voiced vs. voiceless consonants). Therefore it was necessary to create a customized search engine for TELL.

Written in the PERL programming language, the search engine supports regular expression searches, allowing segmental patterns of all kinds to be described. The search engine facilitates the task of the user by having prespecified some metacharacters for natural categories such as consonants, obstruents, vowels, high vowels, etc. Searches can be rendered sensitive to syllable boundaries and stress. Etymological information can be specified as a precondition or can be accessed as a supplement to search results of lexical content. Users can save their search results in a text file for local post-processing. The TELL search engine is accessible from the TELL home page, or directly at http://socrates.berkeley.edu:7037/cgi-

bin/TELLsearch.cgi.

4. Goals for next phase of the TELL project

In the second phase of the TELL project, the database will be expanded in all if the directions that were taken in the first phase, that is, lexically, phonetically, etymologically, and morphologically. Some of the specific goals are listed below:

- Addition of new lexical items such as lowfrequency items and recent borrowings, acquired from selected specialized texts
- Addition of new speaker pronunciations
- Continuation with etymological origin identification
- Continuation with morphological root segmentation

4.1. Expansion of TELL

In addition to completing the tasks which were not fully completed during the first phase, the second phase of the project will see some extensions of the TELL database. These extensions will increase the amount of linguistic data in TELL and will thereby make TELL relevant to a wider audience than before.

The proposed extensions to TELL are the following:

Lexical

- Textual
- Acoustic

They are described in detail in the following sections.

4.1.1. Lexical

This component of the TELL expansion promises to double or triple the size of the TELL database. Currently, TELL currently represents only the headwords from the text sources making up the original master list. Yet both the 2d and 3rd editions of the Turkish-English Oxford dictionaries contain large numbers of subheadwords as well. These items are tagged in the SGML markup of the master texts and easily accessible, yet are not presently included in TELL. It will be a straightforward task to extract the tagged subheadwords and add them to the orthographic MASTER portion of TELL. Users will then be able to search for these items.

A morphologist or a lexical or historical semanticist might, for example, want to examine words containing the reciprocal suffix /-IS/ (whose allomorphs are [-iş, -ış, -uş, -üş]. Though apparently productive, the reciprocal is often semantically opaque. Currently, a search of TELL headwords turns up fewer than 50 reciprocal verbs. Yet the dictionaries actually list hundreds more, embedded within complex entries. Access to these would greatly inform the conclusions reached by the linguist using TELL.

To take another example, the study of Turkish compounds has long been of interest to syntacticians, morphologists and semanticists. The Oxford dictionaries list a great number of compounds, particularly those which are idiomatic or irregular in any way. But the vast majority of these are subheadwords and therefore not represented in the current version of the TELL database. Extracting these items, along with the suffixed and reduplicated subheadwords, would be of the greatest usefulness to users.

4.1.2. Textual: Linking TELL lexemes to a text corpus

The phonologist searching the TELL database is looking for patterns in the data. But what happens when an apparent pattern is shown to admit exceptions? For example, palatal vowel harmony, known to apply generally to suffixes, admits exceptions within roots (e.g. *anne* 'mother', with both front and back vowels). Most words in Turkish have final stress (e.g. *kitáp* 'book', *kitaplár* 'book-pl', *kitap-lar-á* 'book-pl-dat'), but some, e.g. *pencére-ler-e* 'window-pl-dat', do not. Some words in Turkish adhere to Labial Attraction (e.g. *sabun* 'soap'), but others (e.g. *kapı* 'door') do not. Exceptions to a perceived pattern can either mean that the perception is simply illusory – i.e. that the grammar does not enforce that particular pattern at all – or that the pattern *is* enforced by the grammar, but admits exceptions for one reason or another. The linguist deciding between these two outcomes generally considers several factors:

- Is the pattern attested in native items, or is it restricted to loans, therefore most likely an artifact of a pattern holding in the source language?
- Is the pattern restricted to a small number or a large number of items?
- Is the pattern restricted to highly infrequent items or is it found in items of high text frequency as well?

The first and second of these questions can already be answered by the TELL user (see e.g. Inkelas et al. (1998) on Labial Attraction). Etymologies are available for over one third of the items in TELL; the user can restrict searches to the "native" vs. the "nonnative" portions of the vocabulary, or search for loans from any of twenty languages. TELL also reports the number of hits returned for any search, so that the user can evaluate the robustness of any pattern he or she is interested in.

The third bulleted question, however, can be of paramount importance but is not answerable given the information currently in TELL. Information about text frequency (as a rough estimate of frequency in the ambient data available to any given speaker) can be crucial to decisions about the salience of a given pattern for the language learner and in understanding patterns of lexical diffusion and sound change in general. For example, Zimmer (1969) found that in an experimental setting, native speaker listeners preferred nonsense words containing the [a-u] vowel pattern to those containing the (more harmonic) [a-1] vocalism. Zimmer had no explanation for this unexpected effect, but has since speculated (personal communication) that it may have to do with the salience of particular lexical items that include this disharmonic vowel sequence. Text frequency is the only objective measure of salience available. In the second phase of the project, therefore, items in the TELL database will be indexed for frequency of occurrence in a text corpus.

To determine text frequency, selected online and print texts on a variety of topics (principally newspapers and magazines) will be gathered. Print texts will computerized, using the scanning and optical character recognition techniques developed by TELL. The words of these texts will then be checked against TELL or run through Oflazer's morphological parser to determine a root for each word. The frequency with which each word, and each root, appears in the texts will be tallied. Items in the TELL database will then be annotated with the rank order of frequency of the root they contain as well as with the rank order of frequency of the whole word.

The result will greatly illuminate the search results obtained by linguists using TELL. Consider, for example, the phonologist interested in the phenomenon of back-vowel roots which trigger front vowel harmony on suffixes (e.g. *saat* 'hour, watch', which is *saat-i* in the accusative instead of the expected, harmonic *saat-1*), or the phonologist interested in the alleged Labial Attraction pattern (which mandates that high vowels be round if preceded by a labial consonant and the vowel [a]), as occurs in e.g. *sabun* 'soap'.

TELL already informs the researcher that such items are mainly loanwords (in the case of the back vowel roots triggering front harmony, almost exclusively Arabic). It is not unheard of (in fact it is quite common) for linguists to set aside data from borrowings if it violates otherwise true generalizations about a language. In this case, the linguist might (and linguists have) set aside the exceptions to vowel harmony as irrelevant to the grammar of Turkish because they are loans. This approach may be justified for some languages. But Turkish is a language which has long had sustained contact with other languages and which has incorporated and adapted borrowings steadily throughout its history. Borrowings constitute a major part of the lexicon and are completely incorporated into the morphology of the language. Surely for a language like Turkish the relevant question to the synchronic phonologist is not whether an exceptional item like saat 'hour, watch' is borrowed, but whether it is highly frequent and likely to make a major impact on the language learner. That information is currently unavailable. But once it becomes available to the user of TELL, it is likely to revolutionize the linguist's view into the Turkish lexicon. Frequency data provides an indication of the kind of data the learner is likely to be exposed to the most, and therefore to the kind of data most likely to influence the learner's grammar the most heavily. That is the kind of data to which the phonologist will really want to pay the most attention.

A benefit of affiliating TELL with an electronic text corpus is the possibility of extracting

concordances for words in the database. Once the texts are computerized, it is a small matter to extract sentences containing those words so that users can see the words used in context. The texts will not be translated, but anyone acquainted with the language to any degree – student, teacher, native speaker, or educated linguist – will be able to make use of the contextual material in their analysis of the word in question.

4.1.3. Acoustic: linking transcriptions to sound files

The most computationally ambitious and potentially linguistically most exciting goal for the second phase of the TELL project is the association of each phonemic transcription in the database with a sound file so that users may hear the word as it was pronounced by the native speaker consultant. This achievement of this goal will make TELL relevant in new ways to phoneticians and language teachers.

TELL currently has 35 hours of high fidelity recordings (in both digital and analog format) of one of the native speaker's self-elicitation, performed in a sound-protected room at UC Berkeley. These recordings promise to be of inestimable value to phoneticians and phonologists interested in questions such as the following:

- phonetic correlates of stress
- vowel devoicing
- liquid devoicing
- epenthesis into consonant clusters
- final plosive devoicing

• proper characterization of palatal and velar laterals

- nature of /r/ in different phonetic contexts
- degree of release of plosives
- degree of aspiration (voice onset time) of plosives in different phonetic contexts

For example, TELL contains thousands of words with final stress, many of them ending in an open syllable. Users of TELL can use this data to investigate the typologically unusual – perhaps unique – phenomenon of stressed final vowel reduction found in Turkish. TELL will make phonetic data available to users everywhere, permitting even linguists with no other access to a Turkish speaker to do experiments on Turkish.

To take an example, it has been claimed (e.g. Clements & Sezer 1982) that back vowel roots causing front harmony on suffixes end in palatalized consonants. In the case of roots like *saat* 'hour', *saat-i* 'time-acc', the palatalization has

been presumed to be abstract, since Turkish does not have a phonemic contrast between /t/ and $/t^{J}/$ in any other position and impressionistically there is no palatalization on the [t] when word final. TELL, equipped with audio representations of each word, can answer this question. The item saat is recorded in isolation and before vowel-initial and consonantinitial suffixes (as well as in compounds in which it is the first or second member). The linguist interested in the quality of its final consonant can make spectrograms of the numerous tokens of this root and others like it, conducting an experiment over the Internet without having to sit down a single speaker in a lab. The advantages of this for phonetically-oriented linguists without access to native speakers or laboratory facilities cannot be downplayed.

Perhaps most importantly, the user of TELL will no longer be dependent of the transcriptions made by the TELL workers. A linguist who doubts any of the transcriptions in TELL can, with a single keystroke, listen to the same sound file heard by the TELL transcriber and make his or her own judgment. Although TELL transcribers are highly proficient, there are always gray areas in transcription-- where, for example, to draw the line between a noisy consonant-consonant transition and a short epenthetic vowel. Future users of TELL will have access to the source data.

The audio files will also be of use to students and teachers of Turkish. For example, a student who wishes to master the pronunciation of a given word or segment can search for and play aloud that word, or words containing that segment. Teachers wishing to conduct listening comprehension drills can make use of the TELL data to play words at random.

4.1.4. Linking TELL to computational resources at Bilkent

The resources developed by the TELL project complement the computational tools for analyzing Turkish that have been developed at Bilkent University by Prof. Kemal Oflazer. One goal of the second (and final) phase of the TELL project is to unite the two sets of utilities into one super-lexicon of Turkish.

5. References

- Clements, G.N., and E. Sezer, 1982. Vowel and consonant disharmony in Turkish. In H. van der Hulst and N. Smith (eds.), *The Structure of phonological representations, Part II.* Dordrecht: Foris.
- Eyüboglu, I.Z., 1988. Tü*rk Dilinin Etimoloji Sözlüğü*. Istanbul: Sosyal Yayınlar, Birinci Basım.
- Inkelas, S., G. Hansson, A. Küntay, and O. Orgun., 1998. Labial Attraction in Turkish: an empirical perspective. Presented at the International Conference on Turkish Linguistics, Oxford.
- Itô, J. and A. Mester. 1995. Japanese phonology. In Goldsmith, John (ed.), *The handbook phonological theory*. Blackwell.
- Lees, R.B. 1966a. On the interpretation of a Turkish vowel alternation. *Anthropological Linguistics*, 8(9): 32-39.
- Lees, R.B. 1966b. Turkish harmony and the phonological description of assimilation. *Türk dil kurumu yayınları* 255: 279-297.
- Ní Chiosáin, M., and J. Padgett, 1993. Inherent Vplace. *Linguistics Research Center report* #*LRC-93-09*. University of California, Santa Cruz.

- Ozon, M.N., 1962. *Türkçe Yabancı Kelimeler Sözlüğü*. Istanbul: Inkilap ve Aka Kitabevleri, Tan Gazetesi ve Matbaası.
- Ozon, M.N., 1973. *Büyük Osmanlıca Türkçe Sözlük*. Istanbul: İnkilap ve Aka Kitabevleri, Beşinci Basım.
- Püsküllüoğlu, A., 1997. *Türkçedeki Yabancı* Sözcükler Sözlüğü. Ankara: Arkadas Yayınevi.
- Stachowski, S., 1975. *Studien über die Arabischen Lehnworter in Osmanisch-Turkischen*. Wroclaw: Zaklad Narodowy im. Ossolinskich.
- Tzitziles, C., 1987. Griechische Lehnworter im Turkischen: Mit besonderer Berucksichtigung der anatolischen Dialekte. Wien: Verlag der Osterreichischen Akademie der Wissenschaften.
- Zimmer, K., 1969. Psychological correlates of some Turkish morpheme structure conditions. *Language*, 45(2): 309-321.