Creating a database and query-tools for the TELL multi-speaker linguistic corpus *

Jeff Good and Ronald Sprouse
jcgood@socrates.berkeley.edu ronald@uclink.berkeley.edu
1203 Dwinelle Hall #2650
Berkeley, CA 94720-2650

Abstract
The Turkish Electronic Living Lexicon (TELL; http://socrates.berkeley.edu:7037) represents the first large-scale effort to collect transcribed recordings of a large number of partial morphological paradigms (more than 17,000) from several speakers of Turkish. The nature of this data presents a fundamental challenge: how to model the data and design the query tools so that it is possible to find interesting phonological patterns within paradigms and across speakers. The primary dimensions along which paradigms in the database are classified are speaker and lexeme, which together can be used to identify particular utterance sets. We discuss the basic structure of the TELL database and the query tools that have been devised to access it.

0 Introduction
The Turkish Electronic Living Lexicon (TELL; http://socrates.berkeley.edu:7037) (Inkelas et al., 2000) is a detailed database of the Turkish lexicon that is unlike most currently-existing printed and electronic lexical resources. TELL seeks to collect comprehensive information from individual speakers’ idiolects rather than attempting to describe a community norm, as, for example, dictionaries do (though TELL also contains information from two Turkish-English dictionaries and other sources as a subset of its data). To this end, TELL has recorded several native Turkish speakers pronouncing word forms from a large number of partial nominal and verbal inflectional paradigms—more than 17,000 paradigms for each speaker. These words have been transcribed and compiled in a database that provides a means of asking and answering detailed questions about Turkish morphophonology in a way not previously available. Thanks to TELL’s extensive coverage of the Turkish lexicon and detailed information about the speakers’ idiolects, along with TELL’s considerable etymological data, it is possible to rigorously examine claims about Turkish phonological patterns that have been proposed in the literature as well as to uncover new patterns. In addition to this original purpose, we expect TELL to be accessed for a broader range of uses—for instance, students learning Turkish will find it useful to search the database and follow the links to the source audio recordings.

Of comparable existing projects, TELL is probably most similar to Hyperlex (Bird, 1997). Though Hyperlex is a research environment and TELL a language-specific research project, they both have in common that they make audio recordings of a large number of lexical items available to the user. However, they contrast insofar as TELL is not a simple lexicon but a multi-speaker lexicon of paradigms. They also contrast greatly in their user interface. Hyperlex allows for an impressive range of phonological searches. TELL’s search engine is more limited and is specifically designed for ease of use for researchers familiar with Turkish, rather than for maximum flexibility.

TELL is currently in transition. The first version of TELL includes data from one speaker only. The second version will contain data from a second and third speaker, whose data are currently being transcribed. The original digital audio data from these speakers is also being processed for inclusion on the project’s web site (the field recordings of the first speaker are lower quality analog recordings and will not be included). We expect all data from the second speaker to be available in spring 2002, and from the third speaker later in the year. We report in this paper on the second version of TELL currently in development.

* Funding for TELL has been provided by National Science Foundation Awards #SBR-9514355 and #BCS-9911003, Sharon Inkelas, Principal Investigator, and by a Hellman Faculty Fund award at the University of California, Berkeley also awarded to Sharon Inkelas.
From an interface standpoint, the intended endpoint for TELL is the creation of a system wherein the transcribed and recorded data can be searched and examined both within and across individual speakers. Since each of the speakers were asked to produce the same set of paradigms, the database has been designed to allow for searches across speakers but within a particular paradigm.

We refer to the basic data in the database—the paradigms produced by the speakers—as “utterance sets”. These are complex objects containing transcriptions, audio recordings, and tags for a closed set of phonological irregularities. The primary dimensions along which utterance sets are stored and accessed are speaker and lexeme, where lexeme is used as a cover term for an abstract linguistic entity which prompted a speaker to produce a particular utterance set.

In section 1, the basic structure of the data is described, focusing on the relationships between the dimensions of speaker, paradigm, and utterance. In section 2, a broad overview of TELL’s use of an SQL database to solve the representational problems is discussed. Finally, section 3 describes the user interface, presently under development, which has been designed to accommodate the unique properties of the TELL data.

1 The structure of the data

1.1 How the descriptive facts of Turkish have influenced the development of TELL

It is worth mentioning, at first, how some aspects of the structure of Turkish have influenced the development of TELL.

A typical Turkish nominal paradigm consisting of the forms found in the TELL elicitations is given in table 1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Form</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>kitap</td>
<td>‘book’</td>
</tr>
<tr>
<td>Accusative</td>
<td>kitabı</td>
<td>‘book (obj.)’</td>
</tr>
<tr>
<td>First Possessive</td>
<td>kitabım</td>
<td>‘my book’</td>
</tr>
<tr>
<td>Professional</td>
<td>kitapçı</td>
<td>‘bookseller’</td>
</tr>
<tr>
<td>First Predicative</td>
<td>kitabım</td>
<td>‘I am a book’</td>
</tr>
</tbody>
</table>

Table 1: Paradigm for kitap ‘book’

The paradigm in table 1 reveals some important facts about Turkish morphophonology. The first is that it is agglutinative and exclusively suffixing—meaning that parsing words into constituent morphemes is very straightforward. Another important fact is that Turkish syllable structure is fairly limited. Turkish syllables are basically restricted to CVX structure (where X represents vowel length or a consonant). Complex onsets are not permitted, and only a limited range of complex codas are permitted. Stems and suffixes undergo a number of morphophonological alternations that can be uncovered through TELL. One such alternation is in table 1. The stem-final consonant $b/p$ alternates between being voiced and voiceless. This alternation is predictable, however, given the fact that all voiced plosives are devoiced when in coda position in Turkish.

The combination of facts mentioned above means that purely segmental content is largely adequate for storing TELL transcription. The only non-segmental information which needs to be marked for in the transcription is stress, which tends to be word-final but is often found on other syllables. Also, in one of the forms which is elicited for TELL, the first person predicative (kitabım in table 1), stress is predictably on the same syllable as it is in the nominative form.

The fact that Turkish words are relatively easy to parse into morphemes means that we have not had to include morpheme boundaries in our transcriptions. A given word’s stem can be uniquely determined from the word’s nominative form and accusative form. Furthermore, given the stem and a morphological label, any word can be unambiguously parsed into a stem and a suffix.

The fact that Turkish syllable structure is fairly simple has an effect on data storage and queries. Syllable boundaries have not been stored in TELL, as syllabic searches can be implemented using nothing more than one, well-constructed regular expression.

Analyzing Turkish morphology is significantly easier than analyzing many other languages because of the lack of such phenomena as inflection classes, ablaut, or other non-concatenative morphology. The lack of these phenomena in Turkish has allowed for simpler data structures, database implementation, and user interfaces in TELL than might be needed for languages with other kinds of morphological features.

1.2 The information found in a TELL paradigm

Table 2 gives an example of the type of information which is part of a typical TELL paradigm for a
speaker with initials PK.¹

Speaker: PK
Tape: 52BA
Master Id: 10454
Orthography: ka@g@i@d
Meaning: “Paper.”
Source: ox92
Utterances:
Nom: [k@a:i@t]
Acc: [k@a:i@di@]
Poss: [k@a:i@di@m]
Prof: [k@a:i@tc@i@]
Pred: [k@a:i@’di@m]

Table 2: Sample TELL paradigm

The bolded fields in table 2 will be the focus of the discussion since they are the most critical in understanding the overall structure of the data.

The first bolded field, Speaker, simply indicates which speaker produced the set of transcribed utterances. Closely tied to Speaker is Tape, which identifies the tape containing the recorded utterances listed for the paradigm.

The next bolded field, Master Id, is a unique identification number assigned to a putative underlying representation (UR) (not given). By assigning each paradigm a pointer to a master identification number, TELL makes possible to query paradigms that are realizations of the same lexeme across different speakers. Regardless of the particular phonological strings transcribed for the utterances of two speakers, two paradigms are taken to be from the “same” lexeme if they both have the same master identification number. These strings will almost always be very similar between speakers, but they need not be identical. The utterance sets seen in table 4 and table 5 give a case where the utterance sets are clearly from the same lexeme but slightly different for the speakers PK and OO respectively—the difference between the two speakers being the length of the first vowel for some of the utterances in the paradigm.

The final bolded field Utterances contains the actual elicited TELL data. Nominal paradigms contain the nominative (or citation form), preceding the accusative, 1 sg. possessive, professional, and 1 sg. predicative forms. Verbal paradigms contain the infinitive, aorist, and causative forms. These forms were carefully selected because they are especially good for revealing morphophonological alternations.

As indicated in table 2, an utterance is itself a complex object consisting both of a transcription and a grammatical label for the transcription representing its relationship to the other utterances in the paradigm.

1.3 The representational structure of the information

As should be clear from table 2, much of the information for a paradigm is the same no matter which speaker is associated with the set of transcribed utterances. For example, every paradigm with master identification number 10456 represents a word whose UR is something like k@a:id regardless of the complete details of the speaker’s actual pronunciation.

The structure of the data, then, is best understood as a series of disjoint data structures, each containing related information, which can be linked to each other via a master identification number. Tables 3, 4, and 5 give a sense as to how this linking works—where the relevant linking field is the master identification field.

Table 3: Reference information for lexeme 10454

Lexeme: 10454
Orthography: ka@g@i@d
Meaning: “Paper.”
Source: ox92

Table 4: Utterances for PK for lexeme 10454

Lexeme: 10454
Speaker: PK
Tape: 52BA
Utterances:
Nom: [k@a:i@t]
Acc: [k@a:i@di@]
Poss: [k@a:i@di@m]
Prof: [k@a:i@tc@i@]
Pred: [k@a:i@’di@m]

¹The “@” symbol is used with the preceding character to identify non-ASCII letters used in Turkish orthography or phonemic transcription. The precise meaning of “@” varies, depending on the preceding character and the field in which it appears. A full description of “@” appears on the project web site. In addition, TELL uses the apostrophe to mark that a vowel is stressed.

Obviously the representation exemplified by tables 3–5 allows information which is consistent across paradigms to be specified only once. However, more importantly, it also provides a way to link paradigms from different speakers.
As the utterance lists in 4 and 5 show, this sort of abstract linking is important since there is frequently phonological variation between speakers for a given paradigm—here, we see that speaker OO and PK do not pronounce the length for the first vowel of the word ‘paper’ the same way for each morphological form.

1.4 A closer look at the representation of utterance sets

The representation given above for an utterances set, of the type exemplified by tables 4 and 5 was simplified somewhat, and some other pieces of information about utterance sets are important to the organization of TELL. Primary among these are the start time and end time of an utterance on a specified tape. This information is used to allow users to listen to individual elicited forms.

Another sort of information that is indicated in utterance sets are particular phonological “irregularities” found in them. Since the primary goal of TELL is to assist in phonological research, it is helpful to mark particular utterance sets and utterances as exhibiting certain types of phonological phenomena which have been important in research on Turkish phonology. These include such things as irregular stress, disharmony, and unpredictably palatalized consonants.

These irregularities have to be represented on two distinct levels—for the whole utterance set and for one particular utterance. Table 6 illustrates this idea. The whole utterance set represents a stem in Turkish with three phonological irregularities: stem-internal vowel hiatus, an unpredictably-palatal final consonant, and disharmonic vowels. Importantly, the paradigm itself is regular—given the stem-level irregularities, all of the utterances in the paradigm are predictable. Thus, they are not marked as irregular.

A paradigm with irregularities like that in table 6 can be contrasted with the paradigm in table 7. In this case, one utterance in the paradigm shows an irregularity. This particular irregularity, an initial consonant cluster that usually triggers an epenthetic vowel, is noted as an irregularity for that particular form. In addition, the morphological label for the form is marked as being irregular at the paradigm level.

2 The implementation of the database

2.1 Introduction

The basic data structures described in section 1 can be appropriately modelled with a relational database. The particular TELL implementation is a MySQL database which contains three types of tables. The first is a basic master identification table which is used to link reference information with particular utterance sets and to link same-lexeme utterance sets across speakers. The second type of
Table 8: Utterances for OO for lexeme 15378

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Nom: [mu@rettebat]</th>
<th>Acc: [mu@retteba:ti:]</th>
<th>Poss: [mu@retteba:ti@m]</th>
<th>Prof: [mu@rettebatc@i:]</th>
<th>Pred: [mu@retteba:’ti@m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irregularities:</td>
<td>nominative, disharmonic</td>
<td>vowel_shortening</td>
<td>vowel_shortening</td>
<td>vowel_shortening</td>
<td>vowel_shortening</td>
</tr>
<tr>
<td>Start time</td>
<td>672.186</td>
<td>686.226</td>
<td>691.073</td>
<td>675.931</td>
<td>679.797</td>
</tr>
<tr>
<td>End time</td>
<td>673.858</td>
<td>688.269</td>
<td>692.895</td>
<td>677.864</td>
<td>682.29</td>
</tr>
</tbody>
</table>

Table 9: Fragment of lexeme table

<table>
<thead>
<tr>
<th>ID</th>
<th>UR</th>
<th>LXM_TD</th>
<th>ORTH</th>
<th>SRC</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>9</td>
<td>a@ri</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>a@ri@z</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>a@s@ar</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>a@s@i@k</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>a@s@i@kane</td>
<td>ox92</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>a@s@ir</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>a@s@ar</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>a@sim</td>
<td>ox57</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Fragment of source table

As can be seen, the reference information in table 10 includes a pointer (in italics) to the lexeme identification number of each lexeme. This is necessary because, among other reasons, there is not always a one-to-one correspondence between reference entries and individual TELL lexemes.

TELL uses a second reference table for some other types of information. This table contains, among other things, etymological information. Its structure is essentially the same as table 10.

2.4 Speaker utterance tables

The final sort of table employed in the TELL database is the speaker utterance table. At present, data from only two speakers has been transcribed—so, there are only two of these tables. However, the structure is modular and a new speaker can be added without any problem once the data is available. Data from a third speaker will be added to TELL in the near future.

Table 11 gives a fragment of a record from the PK table—that is the table containing all the utterance sets for the speaker with initials PK. This record contains a pointer (in italics) to the relevant master identification number. Table 12 gives a fragment, with different information, for the utterance set, corresponding to the same lexeme as table 11 for speaker OO.

Since the data in tables 11 and 12 is inherently identified as being from a particular speaker, by virtue of the table it is contained in, this information is not specified for each record.

Figure 1 schematizes the TELL database structure. The “lexeme table” serves as the crucial point linking the utterance set tables and the reference tables. In figure 1, the various links between records that is consistent for each lexeme regardless of the speaker. In table 10 is an example of one of two tables used by TELL containing source, orthographic, and gloss information for the lexemes in table 9.
3 The user interface

3.1 Requirements

As is typical of any project like this, there are two broad, potentially conflicting, requirements for the TELL user interface: flexibility and ease of use. We have balanced these two requirements in a range of ways, the most prominent being two types of user interfaces.

The first of these, which we won’t discuss in detail here, is a “dictionary” style search. This is an extremely basic search, with one input field that is searchable only on Turkish orthography. The only other user option is whether the search string matches partial words or whole words only. The output of such a search is a basic dictionary style entry and a link to a digitized recording of the word. This search is intended for users with little or no knowledge of linguistics and minimal knowledge of Turkish. We envision it being the primary search interface for the casual user or for the non-linguist student of Turkish who is curious to hear a word pronounced by a native speaker.

The second search has a configurable advanced interface with a range of features. Importantly, it has a default configuration which is intended to be fairly transparent to a linguist. The user requiring more functionality can then set up additional configuration parameters as desired.

This advanced search interface, with or without special configuration, fulfills all the basic requirements for searching the TELL database. That is, the user can specify search parameters which allow for comparison of paradigms across multiple speakers or just for one speaker.

This functionality is achieved in ways that do not map directly onto the database structure for the simple reason that the database, though computationally fairly intuitive, is not necessarily linguistically intuitive. This fact is most prominent with respect to the master identification number. This number is very useful for linking utterance sets to each other and to reference information. However, it contains no linguistically useful information.

The search interface, thus, does not allow direct reference to the master identification number. Rather, it allows users to search for phonological strings in the TELL transcriptions. Since a string matches some part of an utterance which is directly associated with a paradigm by virtue of its position in the structure of the database, matching on a
transcription can also effectively select a particular paradigm. Thus, the paradigm selection function is primarily handled via phonological searches rather than through selection of master identification numbers.

The specification of searching within a user or across users more straightforwardly maps onto the structure of the database. The user simply specifies which speakers a particular phonological string is supposed to match on.

Thus, with these two fundamental search parameters—phonological string and speaker—the user can isolate utterances, paradigms, and speakers and exploit the most important aspects of the database.

In the following sections, four specific features of the database will be discussed: metacharacters, string searches, syllable template searches, a set of additional possible restrictions on searches, and the search results output format. The first three allow for a large degree of power and flexibility in the searches. The last is important for allowing the user to properly view and access the comparative information in the database—in particular it allows the user to compare paradigms even without being able to search across master identification numbers.

The file located on the TELL web site at /Phil2001/SampleTELLform.html provides a sample version of the TELL search form interface. This sample will be referred to below. It includes two string search fields and two syllable template search fields. Though we don’t discuss the details of it here, registered users of TELL who want more (or fewer) of either type of field can configure the search form accordingly. However, all the basic features of the search form can be seen in the sample form.

3.2 Metacharacters

One important feature of the TELL user interface is its support of metacharacters to search across phonological classes. Such a feature is hardly unique to TELL and can be found in, for example, the Hyperlex system (Bird, 1997).

From the user’s perspective, a metacharacter is simply a capital letter predefined as matching a phonological class. From an implementational standpoint, metacharacters stand in for predefined regular expressions implemented in Perl 5 that search over a range of characters corresponding to some phonological natural class. Some examples of the TELL metacharacters and accompanying Perl regular expressions are given in table 13.

<table>
<thead>
<tr>
<th>Metacharacter</th>
<th>Regular Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>C Consonants</td>
<td>(?:[bcdfghjklmnprstvyz]@?)</td>
</tr>
<tr>
<td>G Voiced stops</td>
<td>(?:<a href="?!@">bcdl</a>[lg]@?)</td>
</tr>
<tr>
<td>K Voiceless stops</td>
<td>(?:[c@][ptk]@?)</td>
</tr>
<tr>
<td>V Vowels</td>
<td>(?:[aeoi]@?)</td>
</tr>
<tr>
<td>I High vowels</td>
<td>(?:[ui]@?)</td>
</tr>
<tr>
<td>R Round vowels</td>
<td>(?:[ou]@?)</td>
</tr>
<tr>
<td>E Front vowels</td>
<td>(?:<a href="?!@">iel</a>[ou]@)</td>
</tr>
<tr>
<td>A Back vowels</td>
<td>(?:<a href="?!@">uao</a>[i@])</td>
</tr>
<tr>
<td>&lt; Word boundary</td>
<td>(?::</td>
</tr>
</tbody>
</table>

Table 13: TELL metacharacters

Before performing a search on the database, the TELL search device simply replaces any occurrence of a metacharacter with the appropriate regular expression. All parentheses are non-capturing in order to avoid interference with back references in user-specified search strings, as will be seen in the next section.

3.3 String searches

The most basic and most powerful way to perform searches on the TELL database is the text string search. In the sample search form at /Phil2001/SampleTELLform.html on the TELL web site, the search fields numbered 1 and 2 correspond to text string searches.

The reason why text string searches offer the most powerful search interface is that they support the full-range of Perl regular expressions. Thus, even though the user has available for use all the TELL metacharacters (of the sort seen in table 13), it is also possible to construct more specific regular expression searches.

An example of the use of regular expression in a TELL search can be seen in the file /Phil2001/SampleTextSearch.html on the TELL web site. The input for two text fields searches for a very small class of words in Turkish: those which end in an unvoiced consonant in the citation (or nominative) form but which underlyingly end in a voiced geminate. To identify these words, it is necessary to enter the schematized search parameters seen in table 14.

| String search 1: | VK< citation |
| String search 2: | V(G)\lI< accusative |

Table 14: Specialized text string search

Table 14 includes two different search parameters. The first, making use of only metacharacters,
limits the search to all words which end in a voiceless stop. The second search parameter makes use of metacharacters and Perl regular expression back-references to specify only searching for accusative forms where the stem ends in a geminate voiced consonant.

In addition to regular expressions, the search schematized in 14 illustrates two other important aspects of the TELL search interface.

The first is that the user specifies which field in the database the search should be applied to—here, the morphological citation (or nominative) and accusative forms were specified. Any other morphological form can be specified, in addition to some of the reference fields, like the orthography field.

The second aspect of this search relates to the discussion of Turkish phonology in section 1.1. It was mentioned there that the overall regularity of the Turkish sound system has allowed TELL to simplify its interface in various ways. A good example is the fact that the second search string in 14 does not specify any sort of stem boundary in order to limit the search to stems with final voiced geminates. Rather, a metacharacter for a high vowel is added to the search string just before the word boundary character. This is because, in Turkish, all consonant-final stems form their accusative with a high vowel suffix.

The TELL advanced interface assumes that the user is aware of such basic information as paradigm formation in Turkish and, thus, includes no special tools to specify such things as, for example, stem boundaries. As mentioned in section 1.1, many languages with less regular morphology probably would require such tools.

An additional parameter of the string search, not discussed above, is that the user can specify which speaker datasets should be searched when trying to match the input string.

3.4 Syllable templates

The text string search discussed in section 3.3 is the most basic and powerful sort of search. A second way to search for phonological strings is by using a syllable template. We expect that this sort of search interface will be comfortable to anyone with basic training in phonology.

In the sample search form at /Phil2001/SampleTELLform.html on the TELL web site, the search fields numbered 1 and 2 correspond to syllable template searches. The basic concept behind this search is to allow the user to enter search parameters into “slots” with phonologically well-defined labels, like onset and coda, rather than having to devise appropriate search strings to perform such a search. The syllable template goes beyond simple segmental parsing and also allows the user to specify such things as the location of word boundaries and whether or not a syllable is stressed.

An example of a possible input to the syllable template is schematized in table 15.

<table>
<thead>
<tr>
<th>Word boundary</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onset</td>
<td>b</td>
</tr>
<tr>
<td>Vowel</td>
<td>a Prosody Stressed</td>
</tr>
<tr>
<td>Coda</td>
<td>i Prosody</td>
</tr>
<tr>
<td>Onset</td>
<td></td>
</tr>
<tr>
<td>Vowel</td>
<td></td>
</tr>
<tr>
<td>Coda</td>
<td></td>
</tr>
<tr>
<td>Word boundary</td>
<td></td>
</tr>
<tr>
<td>Adjacency</td>
<td>Not adjacent</td>
</tr>
</tbody>
</table>

Table 15: Syllable template search

The blank fields in table 15 indicate fields which are merely left unrestricted in the search. This particular search has several specifications. The check in the first Word boundary field means that the first syllable of the template must be the first syllable of the word. This syllable is further specified as having to be stressed and having segmental content ba. Another specification in table 15 also indicates that another syllable in the word must contain the vowel i. The final specification Not adjacent means that the syllable containing i must not appear in the syllable immediately following the first syllable.

The file /Phil2001/SampleSyllSearch.html on the TELL web site shows a search form with the specifications in table 15 inputted in the syllable template labelled 3.

Not included in table 15 is the fact that, just as with the text string search, the user specifies what field in the database the syllabic search should be applied to and which speakers should be included, as well.
3.5 Additional user-specified restrictions

The discussion above has shown how the user can narrow the search down to paradigms with particular phonologies and to particular speakers—this is the most important functionality of the system. However, TELL also offers a few other important parameters for narrowing down the searches.

The first of these is word source. The words elicited for TELL came from a number of sources, most importantly, two editions of the Oxford Turkish-English dictionary and a listing of place names. The user can specify whether the search should be restricted to one or the other of these two types of sources.

The second restriction that can be specified is the etymology of the word. TELL workers have identified etymologies for a large number of Turkish words based on an extensive review of the literature (though TELL recognizes that the origins of particular lexical items thus identified may still be in dispute). For research on Turkish phonology, a word’s etymology is of more than just diachronic interest. The status of Arabic borrowings into Turkish, for example, has been of interest to synchronic phonologists for many years.

A third way that searches can be restricted is by limiting them to particular, phonologically well-defined classes of words. As discussed in section 1.4, TELL identifies and marks both paradigms and forms for particular phonological irregularities. TELL has chosen an important subset of these irregularities and gives the user the option of searching only paradigms showing such irregularities. Examples of the types of irregularities which can be specified are: disharmonic stems (Turkish is famous for being a vowel harmony language), “palatal” stems (stems which unexpectedly trigger front vowel harmony on their suffixes), and Sezer-stress stems (stems with non-final stress adhering to a very particular stress pattern).

The interface for these restrictions is very straightforward. A series of pull-down menus has been placed on the advanced interface page. The user simply chooses the restrictions to be placed on the search. The default value of the menus is for the searches to be unrestricted across the various parameters.

3.6 Results output

The results output is a very important aspect of the interface for more than the obvious reason that it is the only way for the user to see the data in TELL. Since the user cannot actually search for a particular paradigm, but rather must search for a particular form, it is only in the results output that whole paradigms themselves can be viewed.

This is achieved by the fact that, even though matches are made on individual word forms, the results display can show all the utterances in the paradigm where the match was made. For example, in the search schematized in table 14, even though the search itself only specified matches for the citation and accusative forms, the whole nominal paradigm can be displayed.

Furthermore, the results display is also set up so that paradigms from different speakers, corresponding to the same lexeme, will be grouped together. This makes it possible to quickly identify interspeaker variation.

The results output for a multi-speaker search is schematized in 16. A full example of a results output can be found in the file /Phil2001/SampleResults.html on the TELL website. This results output is based on the search entry found at /Phil2001/ResultsSearch.html.

| FIRST PARADIGM | Speaker 1 | utterance set |
| orthography, ref info, etc. | Speaker 2 | utterance set |
| Speaker 3 | utterance set |

| SECOND PARADIGM | Speaker 1 | utterance set |
| orthography, ref info, etc. | Speaker 2 | utterance set |
| Speaker 3 | utterance set |

| LAST PARADIGM | Speaker 1 | utterance set |
| orthography, ref info, etc. | Speaker 2 | utterance set |
| Speaker 3 | utterance set |

Table 16: Schema for results output

The type of results output schematized in table 16, as can be seen, fully incorporates the three important dimensions along which the database is organized: paradigm, speaker, and utterance set.

4 Conclusion

We have described the basic architecture of TELL. This includes the abstract data structures that are part of TELL, the structure of the database itself, and the user interfaces which have been developed to access the database.

What distinguishes TELL from similar linguistic databases is that it contains data in the form of
paradigms across multiple speakers. Ensuring that the database allows for proper storage and access of the data to allow for comparative analysis has been central in the development of the database.

The primary solution within the database for properly linking related data between speakers has been organizing utterance sets into paradigms for each speaker and then, making use of a master identifier number, stating when two paradigms from two different speakers correspond to the same lexeme. This system of master identifiers has also allowed TELL to store reference information for the paradigms in one location rather than with the utterance sets for each speaker, thus avoiding unnecessary redundancy in the database.

The user interface for TELL has been designed to balance flexibility with ease of use. The interface allows for a great flexibility in searching for phonological patterns, including allowing the user to do searches using the full power of Perl 5 regular expressions. Some aspects of the database structure, master identification numbers, for example, cannot be used as parameters in searches. However, by properly structuring the search interface and the results output, it is still possible to fully exploit the comparative aspect of the database.

References
