Design for a Turkish Treebank

Kemal Oflazer[1,3] Dilek Z. Hakkani-Tür[2,3] Gökhan Tür[2,3]

[1]Computing Research Laboratory
New Mexico State University
Las Cruces, NM, USA

[2]Speech Tech. and Res. Laboratory
SRI International
Menlo Park, CA, USA

[3]Dept. of Computer Science
Bilkent University
Ankara, Turkey

{ko,hakkani,tur}@cs.bilkent.edu.tr

ABSTRACT: We present the issues that we have encountered in designing a treebank for Turkish along with rationale for the choices we have made for various representation schemes. In the resulting representation, the information encoded in the complex agglutinative word structures are represented as a sequence of inflectional groups separated by derivational boundaries. The syntactic relations are encoded as labeled dependency relations among segments of lexical items marked by derivation boundaries.

Design for a Turkish Treebank

1 Introduction

In the last decade, treebank corpora such as the Penn Treebank [Marcus *et al.*, 1993] have become a crucial resource for building and evaluating natural language processing tools and applications. Although the compilation of such structurally annotated corpora is time-consuming and expensive, the eventual benefits outweigh this initial cost. With a set of future applications in mind, we have undertaken the design of a treebank corpus architecture for Turkish, which we believe encodes the lexical and structural information relevant to Turkish. In this paper, we present the issues that we have encountered in designing a treebank for Turkish along with rationale for the representation choices we have made. In the resulting representation, the information encoded in the complex agglutinative word structures are represented as a sequence of inflectional groups separated by derivational boundaries. A tagset reduction is not attempted as any such reduction leads to removal of potentially useful syntactic markers, especially in the encoding of derived forms. At the syntactic level, we have opted to just represent relationships between lexical items (or rather, inflectional groups) as dependency relations. The representation is extensible so that relations between lexical items can be further refined by augmenting syntactic relations by finer distinctions which are more semantic in nature.

2 Turkish: Morphology and Syntax

Turkish is an Ural-Altaic language, having agglutinative word structures with productive inflectional and derivational processes. Derivational phenomena have rarely been addressed in designing tagsets and in the context of Turkish, this may pose challenging issues for Turkish, as the number of forms one can derive from a root form may be in the millions [Hankamer, 1989].

Turkish word forms consist of morphemes concatenated to a root morpheme or to other morphemes, much like beads on a string. Except for a few exceptional cases, the surface realizations of the morphemes are conditioned by various morphophonemic processes such as vowel harmony, vowel and consonant elisions. The morphotactics of morpheme sequencing is quite complex and as a result of productive derivational phenomena, word structures may involve several derivations. For instance, the derived determiner `sağlamlaştırdığımızdaki`[1] would be represented as:[2]

```
sağlam+Adj^DB+Verb+Become^DB+Verb+Caus+Pos^DB+Adj+PastPart+P1sg^DB
    +Noun+Zero+A3sg+Pnon+Loc^DB+Det
```

Marking such a word as a determiner and ignoring anything that comes before the last part of speech would ignore the fact that the stem is an adjective which may have syntactic relations with preceding words such as an adverbial modifier, or that there is an intermediate causative (hence transitive) verb which may have an object NP to its left.

A recent experiment that we conducted on about 250,000 Turkish words in news text revealed

[1] Literally, "(existing) at the time we caused (something) to become strong". Obviously this is not a word that one would use everyday. Turkish words found in typical text average about 10 letters.

[2] The morphological features other than the obvious POSs are: `+Become`: become verb,`+Caus`: causative verb, `+PastPart`:Derived past participle, `+P1sg`: 1sg possessive agreement, `+A3sg`: 3sg number-person agreement,`+Zero`:Zero derivation with no overt morpheme, `+Pnon`: No possessive agreement, `+Loc`:Locative case, `+Pos`: Positive Polarity. `^DB` denotes a derivation boundary. A comprehensive list of morphological features is given in Appendix A.

that there were over 6000 distinct morphological feature combinations when root morphemes were ignored. Although this is less than the much larger numbers quoted by Hankamer who considered the generative capacity of the derivations, it is nevertheless much larger than the distinctions encoded by the tagsets of languages like English or French. But what is important is not the size of the potential tagset but rather (i) the fact that there is no a priori limit on it as the next set of million words that one looks at may contain another 3000 distinct feature combinations, and (ii) the nature of the derivational information.

On the syntax side, although Turkish has unmarked constituent order that is SOV, it is considered a free-constituent order language as all constituents including the verb, can move freely as demanded by the discourse context with very few syntactic constraints [Erguvanlı, 1979]. Case marking on nominal constituents usually indicates their syntactic role. Constituent order in embedded clauses is substantially more constrained but deviations from the default order can be found. Turkish is also a pro-drop language as the subject, if necessary, can be elided and recovered from the agreement markers on the verb. Within noun phrases, there is a loose order with specifiers preceding modifiers, but within each group, order (e.g., between cardinal and attributive modifiers) is mainly determined by which aspect is to be emphasized. For instance the Turkish equivalents of *two young men* and *young two men* are both possible: the former being the neutral case or the case where youth is emphasized, while the latter is the case where the cardinality if emphasized. A further but relatively minor complication is that various verbal adjuncts may intervene in well-defined positions within NPs causing discontinuous constituents.

3 What information needs to be represented?

We expect this treebank to be used by a wide variety "consumers", ranging from linguists investigating morphological structure and distributions, syntactic structure, constituent order variations to computational linguist extracting language models or evaluating parsers, etc. We would therefore employ an extendable multi-tier representation, so that any future extensions can be easily incorporated if necessary.

3.1 Representing Morphological Information

At the lowest level we would like to represent three main aspects of a lexical item:

- The word itself, e.g., `evdekiler`
- The lexical structure, as a sequence of free and bound morphemes (including any elided morphophonological material and meta symbols for relevant phonological categories), e.g., `ev+DA+ki+lAr` (where for instance `D` represents a set of dental consonants, `H` a set of high-vowels and `A` represents non-round front vowels.)
- The morphological features encoded by the word as a sequence of morphological and POS feature values all of which except the root are symbolic, e.g., `ev+Noun+A3sg+Pnon+Loc-^DB+Det^DB+Noun+Zero+A3pl+Pnon+Nom`. A point to note about this representation is that, information that is conveyed covertly by zero-morphemes that are not explicit in the lexical representation, is represented here. (e.g., if a plural marker is not present then the noun is singular hence `+A3sg` is the feature supplied even though there is no overt morpheme.) A comprehensive list of morphological feature symbols is given in Appendix A.

The first two components of the morphological information do not deserve any more details for the purposes of this presentation. The third component with its relation to lexical tag information needs to be detailed further.

The prevalence of productive derivational word forms bring a challenge to representing such information using a finite (and possibly reduced) tagset. The usual approaches to tagset design, typically assume that the morphological information associated with a word form can be encoded using a finite number of cryptically coded symbols from some set whose sizes range from few tens (e.g., Penn Treebank tag set [Marcus *et al.*, 1993]) to hundreds or even thousands (e.g., Prague Treebank tagset, [Hajič, 1998]). But such a finite tagset approach for languages like Turkish inevitably leads to loss of information. The reason for this is that the morphological features of intermediate derivations can contain markers for syntactic relationships. For instance in the example we gave earlier `sağlamlaştırdığımızdaki`, it is the intermediate verbal and nominal derivations and the related inflectional features that link this word to any subject and object NPs (of the verbal part) and verbal adjuncts to the left. Leaving out this information within a fixed-tagset scheme may prevent crucial syntactic information from being represented.

For these reasons we have decided not to compress in any way the morphological information associated with Turkish words and represent such words as a sequence of *inflectional groups* (IGs hereafter), separated by `^DB`s denoting derivation boundaries, in the following general form:

$$\texttt{root+Infl}_1\texttt{\^{}DB+Infl}_2\texttt{\^{}DB+}\cdots\texttt{\^{}DB+Infl}_n$$

where $\texttt{Infl}_i$ denote relevant inflectional features including the part-of-speech for the root or any of the derived forms. For instance, the derived determiner `sağlamlaştırdığımızdaki` would be represented by the 6 IGs:

1. `sağlam+Adj`
2. `+Verb+Become`
3. `+Verb+Caus+Pos`
4. `+Adj+PastPart+P1sg`
5. `+Noun+Zero+A3sg+Pnon+Loc`
6. `+Det`

Note that it is possible to come up with a finite inventory of IGs which can be compactly coded, but we feel that apart from saving storage such an encoding serves no real purpose while the resulting opaqueness prevents facilitated access to component features.

Turkish is also very rich in lexicalized and non-lexicalized collocations The lexicalized collocations are much like what one would find in other languages. On the other hand, non-lexicalized collocations can be divided into two groups. In the first group, we have compound and support verb formations where there are two or more lexical items the last of which is a verb. Even though the other components can themselves be inflected, they can be assumed to be fixed for the purposes of the collocation and the collocation assumes its inflectional features from the inflectional features of the last verb which itself may undergo any morphological derivation or inflection process. For instance, the idiomatic verb *kafayı çek-* `(kafa+Noun+A3sg+Pnon+Acc çek+Verb+...)` (literally to pull the head) means *to get drunk*, and these two tokens essentially behave together as far as syntax goes. The second group of non-lexicalized collocations involve full or partial duplication of verb forms or morphemes. For instance, the aorist marked verb sequence *gelir gelmez* `(gel+Verb+Pos+Aor+A3sg gel+Verb+Neg+Aor+A3sg)` actually functions as a temporal adverbial meaning *as soon as ... comes*. Note that these formations (usually involving word reduplications of the sort $\omega\ \omega$ or $\omega\ x\ \omega$) are beyond the formal power of finite state mechanisms hence are not dealt within a finite state morphological analyzer.

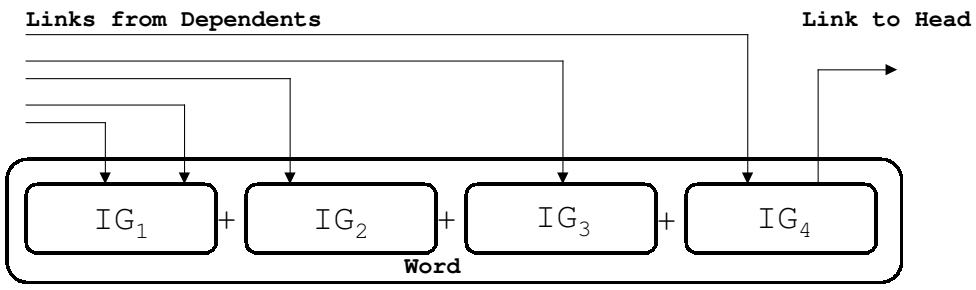


Figure 1: Links and Inflectional Groups

3.2 Representing Syntactic Relations

We would like to represent the syntactic relations between lexical items (actually between inflectional groups as we will see in a moment) using a simple dependency framework. Our arguments for this choice essentially parallels those of recent works on this topic [Hajič, 1998, Skut *et al.*, 1997, Lepage *et al.*, 1998]. Free constituent ordering and discontinuous phrases make the use constituent-based representations rather difficult and unnatural to employ. It is however possible to use constituency where it makes sense and bracket sequences of tokens to mark segments in the texts whose internal dependency structure would be of little interest. For instance time–date expressions or multiword proper names are two examples which can be bracketed a priori as chunks and then related to other constituents. If necessary, any constituent-based representation can be extracted from the dependency representation [Lin, 1995].

An interesting observation that we can make about Turkish is that, when a word is considered as a sequence of IGs, syntactic relation links only emanate from the last IG of a (dependent) word, and land on one of the IGs of the (head) word on the right (with minor exceptions), as exemplified in Figure 1. A second observation is that, with minor exceptions, the dependency links between the IGs, when drawn above the IG sequence, do not cross (although this is not a concern here). Figure 2 part a), shows a dependency tree for the following sentence laid on top of the words segmented along IG boundaries.

(1) Bu eski bahçe-de+ki gül-ün
`bu(this)+Det eski(old)+Adj bahçe(garden)+A3sg+Pnon+Locˆ+Det gül(rose)+Noun+A3sg+Pnon+Gen`
The growth of the rose

böyle büyü+me-si
`böyle(like-this)+Adv büyü(grow)+Verb+Posˆ DB+Noun+Inf+A3sg+P3sg+Nom`
like this in this old garden impressed everybody.

herkes-i çok etkile-di.
`herkes(everybody)+Pron+A3sg+Pnon+Acc çok(very)+Adv etkile(impress)+Verb+Pos+Past+A3sg`

The syntactic relations that we have currently opted to encode in our syntactic representation are the following:

1. Subject	2. Object,	3. Modifier (adverbs/adjectives)
4. Possessor,	5. Classifier	6. Determiner
7. Dative Adjunct	8. Locative Adjunct	9. Ablative Adjunct
10. Instrumental Adjunct		

Some of the relations above perhaps require some more clarification. *Object* is used to mark objects of verbs and the nominal arguments of postpositions. A *classifier* is a nominal modifier in nominative case (as in *book cover*) while a *possessor* is a genitive case-marked nominal modifier. For verbal adjuncts we indicate the syntactic relation with a marker paralleling the case marking though the semantic relation they encode is not only determined by the case marking but also the lexical semantics of the head noun and the verb they are attached to. For instance a dative adjunct can be a *goal*, a *destination*, a *beneficiary* or a *value carrier* in a transaction, or a *theme*, while an ablative adjunct may be *reason* a *source* or a *theme*. Although we do not envision the use of such detailed relation labels at the outset, such distinctions can be certainly be useful in training case-frame based transfer modules in machine translation systems to select the appropriate prepositions in English for instance.

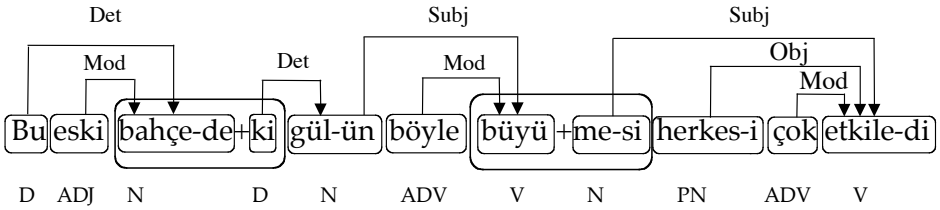
3.3 Example of a Treebank Sentence

In this section we present the detailed representation of a Turkish sentence in the treebank. Each sentence is represented by a sequence of attribute lists of the words involved, bracketed with tags `<S>` and `</S>`.[3] Figure 2 shows the treebank encoding for the sentence given earlier. Each word is bracketed by `<W>` and `</W>` tags. The `IX` denotes the number or index of the word. `LEM` denotes the lemma of the word, as one would find in a dictionary. For verbs, this would be typically be an infinitive form, while for other word classes it would usually be the root word itself. `MORPH` indicates the morphological structure of the word as a sequence of morphemes, essentially corresponding to the lexical form. The morphemes may involve meta-symbols (mentioned earlier) for indicating any phonological classes of symbols. `IG` is a list of pairs of an integer and an inflection group. `REL` encodes the relationship of this word, as indicated by its last inflection group, to an inflectional group of some other word. The first component of `REL` is the index of a word, the second component is the number of the inflection group in that word that is this word's last inflection is linked to, and the third component is a list of relation labels. For example, the 4th and 5th words in the sentence are subject and and adverbial modifier, respectively, of the verb in the first IG of the 6th word, while the 2nd IG of the same word (6) is the subject of the main verb of the word 9. We have only used simple syntactic relation names in the example but more certainly can be added. For instance modifiers can be further classified into attributive, cardinal, etc., while the object may further be marked as theme or patient, as discussed earlier.

4 Conclusions and Future Work

As we mentioned at the outset, our current work has concentrated on resolving the issues in encoding Turkish treebanks. Our next step will involve developing an editing tool for visualizing and editing the treebank corpus in both textual and graphical forms, and integrating tools that we have already developed for tokenizing, morphological analysis, collocation processing and morphological disambiguation. We also expect to integrate a dependency parser to generate dependency parses

[3] Words in this context may actually be a lexicalized or non-lexicalized collocations.



Last line shows the final POS for each word.

a) Dependency structure for a sample Turkish Sentence

```
<S>
<W IX=1 LEM="bu" MORPH="bu" IG=[(1, "bu+Det")] REL=[(3,1,(DETERMINER)]> Bu </W>

<W IX=2 LEM="eski"' MORPH="eski" IG=[(1, "eski+Adj")] REL=[3,1,(MODIFIER)]> eski> </W>

<W IX=3 LEM="bahçe" MORPH="bahçe+DA+ki" IG=[(1, "bahçe+A3sg+Pnon+Loc") (2, "+Det")]
REL=[4,1,(DETERMINER)]> bahçedeki </W>

<W IX=4 LEM="gül" MORPH="gül+nHn" IG=[(1,"gül+Noun+A3sg+Pnon+Gen")] REL=[6,1,(SUBJECT)]>
gülün </W>

<W IX=5 LEM="böyle" MORPH="böyle" IG=[(1,"böyle+Adv")] REL=[6,1,(MODIFIER)]> böyle </W>

<W IX=6 LEM="büyümek" MORPH="büyü+mA+sH" IG=[(1,"büyü+Verb+Pos") (2,
"+Noun+Inf+A3sg+P3sg+Nom")] REL=[9,1,(SUBJECT)]> büyümesi </W>

<W IX=7 LEM="herkes" MORPH="herkes+yH" IG=[(1,"herkes+Pron+A3sg+Pnon+Acc")]
REL=[9,1,(OBJECT)]> herkesi </W>

<W IX=8 LEM="çok" MORPH="çok" IG=[(1,"çok+Adv'')] REL=[9,1,(MODIFIER)]> çok </W>

<W IX=9 LEM="etkilemek" MORPH="etkile+DH" IG=[(1, "etkile+Verb+Pos+Past+A3sg")] REL=[]>
etkiledi </W>

</S>
```

b) Treebank Encoding

Figure 2: Encoding a Turkish sentence in the treebank

wherever possible and have a human operator disambiguate the parses. Even though we now have access to millions of word of morphologically analyzed Turkish text, we can only start creating the Turkish treebank after having built such a tool.

References

[Erguvanlı, 1979] Eser Erguvanlı. *The Function of Word order in Turkish.* PhD thesis, University of California, Los Angeles, 1979.

[Hajič, 1998] Jan Hajič. Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajicova, editor, *Issues in Valency and Meaning: Studies in Honour of Jarmila Panenova.* Karolinum – Charles University Press, Prague, April 1998.

[Hankamer, 1989] Jorge Hankamer. Morphological parsing and the lexicon. In W. Marslen-Wilson, editor, *Lexical Representation and Process.* MIT Press, 1989.

[Lepage *et al.*, 1998] Yves Lepage, Ando Shin-Ichi, Akamine Susumu, and Iida Hitoshi. An annotated corpus in japanese using tesnire's structural syntax. In *Proceedings of COLING-ACL'98 Workshop on the Processing of Dependency-based Grammars*, 1998.

[Lin, 1995] Dekang Lin. A dependency-based method for evaluation broad-coverage parsers. In *Proceedings of IJCAI'95*, 1995.

[Marcus *et al.*, 1993] Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewitz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 1993.

[Skut *et al.*, 1997] Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of Fifth Conference on Applied Natural Language Processing*, 1997.

A Turkish Morphological Features

- **Major Parts of Speech**: `+Noun, +Adj, +Adv, +Conj, +Det, +Dup, +Interj, +Ques, +Verb, +Postp, +Num, +Pron, +Punc`. (`+Dup` category contains onomatopoeia words which only appear as duplications in a sentence.)
- **Minor Parts of Speech**: These typically follow a major POS to further subdivide that class, or to indicate the kind of derivation involved.
  - `+Card, +Ord, +Percent, +Range, +Real, +Ratio, +Distrib, +Time` after `+Num`
  - `+Inf, +PastPart, +FutPart, +Prop, +Zero` after `+Noun`
  - `+PastPart, +FutPart, +PresPart`, after `+Adj`
  - `+DemonsP, +QuesP, +ReflexP, +PersP, +QuantP` after `+Pron`
- The following (mostly semantic) markers are used after derivations to indicate the kind of derivation involved:
  - After `+Adv` derived from verbs: `+AfterDoingSo, +SinceDoingSo, +As (he does it), +When, +ByDoingSo, +While, +AsIf, +WithoutHavingDoneSo`.

- After `+Adv` derived from Adjectives: `+Ly` (equivalent to the English *+ly* derivation.)
- After `+Adv` derived from temporal nouns: `+Since`
- After `+Adj` derived from nouns: `+With, +Without +SuitableFor, +InBetween, +Rel`.
- After `+Noun` derived from adjectives: `+Ness` (as in red vs. redness)
- After `+Noun` derived from nouns: `+Agt` (someone involved in some way with the stem noun), `+Dim` (Diminutive),
- After `+Verb` derived from nouns or adjectives: `+Become` (to become like the noun or adjective in the stem) `+Acquire` (to acquire the noun in the stem)
- A `+Zero` appears after a zero morpheme derivation.

- Nominal forms (Nouns, Derived Nouns, Pronouns, Participles and Infinitives) get the following additional inflectional markers:
  1. Number/Person Agreement: `+A1sg, +A2sg, +A3sg, +A1pl, +A2pl, +A3pl`.
  2. Possessive Agreement: `+P1sg, +P2sg, +P3sg, +P1pl, +P2pl, +P3pl, +Pnon` (no overt agreement).
  3. Case: `+Nom, +Acc, +Dat, +Abl, +Loc, +Gen, +Ins`.
- Adjectives (lexical or derived) do not take any inflection, except `+Adj+PastPart` and `+Adj+FutPart` will have a `+Pxxx` (possessive agreement as above) to mark verbal agreement. Any other inflection to adjectives implies type-raising to nouns and the inflection goes onto the noun after a 0-morpheme derivation.
- Verbs have 2 sets of markers which are treated as derivations:
  1. Voice: `+Verb+Pass, +Verb+Caus, +Verb+Reflex +Verb+Recip`. A verb form may have multiple causative markers.
  2. Compounding/Modality: `+Verb+Able` (able to verb), `+Verb+Repeat` (verb repeatedly), `+Verb+Hastily` (verb hastily), `+Verb+EverSince` (have been verb-ing ever since), `+Verb+Almost` (almost verb-ed but did not), `+Verb+Stay` (stayed frozen while verb-ing), `+Verb+Start` (start verb-ing immediately)
- Verbs also get the following inflectional markers:
  1. Polarity: `+Pos, +Neg`
  2. Tense - Aspect - Mood: A finite verb may have 1 or 2 of `+Past` (past tense) `+Narr`, (narrative past tense) `+Fut`, (future tense) `+Aor`, (Aorist, may indicate habitual, present, future, you name it) `+Pres` (present tense, for predicative nominals or adjectives) `+Desr` (desire/wish) `+Cond` (conditional) `+Neces` (Necessitative, must) `+Opt` (optative, let me/him/her verb), `+Imp` (imperative), `+Prog1` (Present continuous, process), `+Prog2` (Present continuous, state).
  3. Verbs also have number person agreement markers and an optional copula.