Morpheme Segmentation in the METU-Sabancı Turkish Treebank

Ruket Çakıcı
Computer Engineering Department
Middle East Technical University
Ankara, Turkey
ruken@ceng.metu.edu.tr

Abstract

Morphological segmentation data for the METU-Sabancı Turkish Treebank is provided in this paper. The generalized lexical forms of the morphemes which the treebank previously lacked are added to the treebank. This data maybe used to train POS-taggers that use stemmer outputs to map these lexical forms to morphological tags.

1 Introduction

METU-Sabancı Treebank is a dependency treebank of about 5600 modern day Turkish sentences annotated with surface dependency graphs (Atalay et al., 2003; Oflazer et al., 2003). The words in the treebank are annotated with their morphological structure. However, only the tag information is used in the annotations. These tags are combined to create what was called inflectional groups (IG). An IG field contains one or more inflectional morpheme tag groups separated by derivational boundaries. An example IG with two inflectional groups from Figure 1 is *IG='[(1,"dayan+Verb+Pos")(2,"Adv+AfterDoingSo")]'*. A derivational boundary marking a part-of-speech change (from Verb in the first IG to Adverb in the second IG) is seen here.

The lexical forms of the morphemes and the lemma information were initially planned to be included in the annotated data. Thus the annotation files have fields MORPH and LEM that are empty in the current version. With this study, we aim to include this missing information and provide the treebank data in a more complete form for further studies. The sentence in (1) is taken from the treebank and is shown with the intended representation given in Figure 1. The LEM field contains the lemma information whereas the MORPH field contains the lexical representations of the morphemes involved in forming the word. For the explanations of the rest of the fields the reader is referred to Atalay et al. (2003) and Oflazer et al. (2003).

(1)

Kapının	kenarındaki	duvara	dayanıp
door	**side**	**wall**	**lean**
bize	baktı	bir	an.
us	**looked**	**one**	**moment**

(He) looked at us leaning on the wall next to the door, for a moment.

Part-of-speech (POS) tagging with simple tags such as *Verb*, *Adverb* etc. is not appropiate and sufficient for agglutinative languages like Turkish. This is especially obvious in the Turkish dependency treebank. A derived word may have arguments (dependents) of its root but it may have different dependencies regarding its role in the sentence. Most of the voice changes, relativisation and other syntactic phenomena is handled through morphology in Turkish (Çakıcı, 2008). Therefore morphological taggers for agglutinative languages are usually preferred over simple part-of-speech taggers since there is not a simple part-of-speech tagset for Turkish. METU-Sabancı treebank is the only available syntactically annotated data for Turkish. Providing the morphological segmentation of the words in the treebank will make it easier to map the morphological structure in the IG fields to the wordforms.

Proceedings of the 6th Linguistic Annotation Workshop, pages 144–148,
Jeju, Republic of Korea, 12-13 July 2012.

<S No="3">

<W IX="1" LEM="kapı" MORPH="kapı+nHn" IG='[(1,"kapı+Noun+A3sg+Pnon+Gen")]' REL="[2,1,(POSSESSOR)]"> Kapının </W>

<W IX="2" LEM="kenar" MORPH="kenar+nHn+DA+ki" IG='[(1,"kenar+Noun+A3sg+P3sg+Loc")(2,"Adj+Rel")]' REL="[3,1,(MODIFIER)]"> kenarındaki </W>

<W IX="3" LEM="duvar"MORPH="duvar+yA" IG='[(1,"duvar+Noun+A3sg+Pnon+Dat")]' REL="[4,1,(OBJECT)]">duvara </W>

<W IX="4" LEM="dayanmak" MORPH="dayan+Hp" IG='[(1,"dayan+Verb+Pos")(2,"Adv+AfterDoingSo")]' REL="[6,1,(MODIFIER)]"> dayanıp </W >

<W IX="5" LEM="bize" MORPH="biz+yA" IG='[(1,"biz+Pron+PersP +A1pl+Pnon+Dat")]' REL="[6,1,(OBJECT)]"> bize </W>

<W IX="6" LEM="bakmak" MORPH="bak+DH" IG='[(1,"bak+Verb+Pos +Past+A3sg")]' REL="[9,1,(SENTENCE)]">baktı </W>

<W IX="7" LEM="bir" MORPH="bir" IG='[(1,"bir+Det")]' REL="[8,1,(DETERMINER)]"> bir </W>

<W IX="8" LEM="an" MORPH="an" IG='[(1,"an+Noun+A3sg+Pnon+Nom")]' REL="[6,1,(MODIFIER)]"> an </W>

<W IX="9" LEM="." MORPH="." IG='[(1,".+Punc")]' REL="[,()]"> . </W>

</S>

Figure 1: The encoding of the sentence in (1) in the dependency treebank

The segmentation data provided here is universal unlike the tag mapping in IGs, thus it may also be applied to morphological information decodings in alternative formats which may prove more useful for parsing Turkish dependency treebank sentences with structures other than the one in use at the moment.

The example in (2) shows a not-so-complicated Turkish word from the treebank *düşünmediklerim – the ones that I did not think of*. The lexical segmentation of this word is as shown in (2b), and the corresponding morpheme functions are shown with the tags in (2c). Here, *Neg* represents the negative morpheme for verbs, *Rel* represents the nominalization morpheme that is also used for relative clause formation in Turkish (PastPart in d) and *Agr1sg* is used for aggreement (Poss1sg in d). (2d) shows the IG field for this word in the treebank.

(2) a). düşünmediklerim
b). düşün+me+dik+ler+im
c). think+Neg+Rel+Plural+Agr1sg
d). (1, "düşün+Verb+Neg")
(2,"Noun+PastPart+Plu+Poss1sg+Nom")

The MORPH information to be added in the case of (2) will be *düşün+mA+dHk+lAr+Hm*. Generalization is aimed when adding this information to the treebank. Therefore we will not use the surface realizations or allomorphs as in (2b) but the lexical forms of the morphemes instead. The meaning of the capital letters in these lexical forms are given in Section 2.

There are approximately 60000 words in the treebank. Reliable POS tagging requires morphological analysis and disambiguation of the words used. However, a full part of speech tagger that assigns morphological structures like the ones adopted in the treebank is not currently available freely. The reason for that partly is the fact that the tag information in the treebank is too long and this causes sparse data problems when training classifiers with the full tag sequences as in (2d). The morphological tags include all kinds of derivational and inflectional morphemes. Moreover, they include some tags that do not correspond to any surface form such as the *Nom* tag in (2d). We believe morphological segmentation information included will make training and developing POS taggers for the Turkish treebank possible by providing the mapping between the lexical/surface morphemes/allomorphs to the tags or tag groups in the treebank data.

In the next section the lexical forms of the morphemes are described and are related to the data in the treebank. In Section 3 a brief history of part-of-speech tagging in Turkish is covered. The annotation method is then described in Section 4 and conclusion and future work section follows.

2 The Morpheme Set and the Mapping

Oflazer et al. (1994) give a list of all the morphemes in Turkish morpheme dictionary. These contain some compositional derivational morphemes as well. What we mean by that is that the derivation is productive and the semantics of it can be guessed with compositional semantics principles. Moreover, most morphosyntactic phenomena such as relativization and voice changes are marked on the verb as derivational morphology in the Turkish treebank.

Case	+DA, +nHn, +yA, +DAn, yH, ylA, +nA, +nH, +ndA, +ndAn
Agreement	+lAr, +sH, +m, +n, +lArH, +mHz, +nHz
Person	+sHnHz, +yHm, +sHn,+yHz,+sHnHz,+lAr, 0, +m, +n, +k,+nHz +z, +zsHn, +zsHnHz, +zlAr
Voice	+Hş, +n, +Hl, +DHr, +t, +Hr,
Possessive	+sH, +lArH, +Hm, +Hn,+HmHz, +HnHz
Derivation	+cA, +lHk, +cH, +cHk, +lAş, +lA, +lAn, +lH, +sHz, +cAsHnA, +yken, +yArAk, +yAdur, +yHver, +Akal, +yHver, +yAgel, +yAgör ,+yAbil+, yAyaz, +yAkoy, +yHp, +yAlH, +DHkçA, +yHncA, +yHcH, +mAksHzHn, +mAdAn, +yHş, +mAzlHk
Rel/Nom	+ki, +yAn, +AsH, +mAz, +dHk, +AcAK, +mA, +mAk
Tense	+ydH, +ysA, +DH,+ymHş, +yAcAk, +yor, +mAktA, +Hr
Negative	+mA, +yAmA
Mood	+yA, +sA, +mAlH, 0(imperative)

Table 1: Morpheme list

The list of morphemes in Oflazer et al. (1994) is given in Table 1. The capital letters in the lexical forms of these morphemes represent generalization over allomorphs of the morpheme. *H* in the morpheme representations designates a high vowel (*i,ı, u, ü*) whereas *D* can be instantiated as one of *d,t* and *A* as one of *a,e*. These abstactions are necessary for representing the allomorphs of these morphemes in the lexical forms in a compact manner. The surface representations for the morphemes conform to certain voice changes such as vowel harmony present in Turkish and these capital letters are instantiated as one of the surface letters they represent.

Some morphemes in the list are shown as 0 such as the 3rd person singular. This means that these morphemes are not realized in the surface form. Moreover, some morphemes are ambiguous in the surface form and, furthermore, in grammatical functions such as *+AcAk*, the future tense morpheme and *+AcAk* , the relativization morpheme. Another example to this is *+lAr*, the plural marker of nominal morphology and the third person plural marker in verbal morphology. Agreement class contains the plural marker *+lAr* and also the agreement morphemes attached to nominalizations and relativization. We have separated these in this list because of their functional/grammatical differences with the possessive markers on nouns although they have the same lexical and surface forms.

In this study, we use the two modes of the Turkish morphological analyser built for the Turkish dependency treebank (Atalay et al., 2003) using Xerox Research Centre Finite State Toolkit (Karttunen and Beesley, 2003). The *lexmorph* mode creates morphological tag analyses similar to IGs used in the treebank and the *lexical* mode creates the generalized lexical forms consisting of the morphemes in Table 1.

A1pl	NotState	A1sg	Noun
A2pl	Num	A2sg	Opt
A3pl	Ord	A3sg	P1pl
Abl	P1sg	Able	P2pl
Acc	P2sg	Acquire	P3pl
Adj	P3sg	Adv	Pass
Agt	Past	AfterDoingSo	PastPart
Aor	PCAbl	As	PCAcc
AsIf	PCDat	Become	PCGen
ByDoingSo	PCIns	Card	PCNom
Caus	PersP	Cond	Pnon
Conj	Pos	Cop	Postp
Dat	Pres	Demons	PresPart
DemonsP	Prog1	Desr	Prog2
Det	Pron	Distrib	Prop
Dup	Punc	Equ	Ques
FitFor	QuesP	Fut	Range
FutPart	Real	Gen	Recip
Hastily	Reflex	Imp	ReflexP
InBetween	Rel	Inf	Related
Ins	Since	SinceDoingSo	Interj
JustLike	Stay	Loc	Time
Ly	Verb	Nar	When
Neces	While	Neg	With
Without	Ness	WithoutHavingDoneSo	Nom
Zero			

Table 2: Morphological tags in the METU-Sabancı Turkish treebank data.

3 Morphological tagging of Turkish

The first attempt in automatically recognizing Turkish morphology is a two-level system of finite state transducers. Oflazer (1994) implements the morphotactic rules of Turkish that are explained in Oflazer et al. (1994) by using PC-KIMMO which is a two level morphological analyser system developed by Antworth (1990). A Xerox FST implementation of this morphological analyser was also used for morphological analysis in METU-Sabancı Treebank (Atalay et al., 2003; Oflazer et al., 2003).

When the level of morphological ambiguity is considered in Turkish, morphological disambiguators that choose between different analyses are vital for practical NLP systems with a morphological processing component.Oflazer and Tür (1996) and Oflazer and Tür (1997) are two of the early disambiguators that use hybrid models of hand crafted rules and voting constraints modelling the context of the word to be tagged. A purely statistical model is created by Hakkani-Tür et al. (2002).

Yüret and Türe (2006) use decision trees and train a separate model for each of the morphological features/tags the morphological analyser creates. These features are the 126 morphological tags that Oflazer (1994)'s morphological analyser creates. They report a tagging result of 96% when a separate classifier is trained for each tag and 91% when decision lists are used to tag the data without the help of a morphological analyser. The training data was a semi-automatically disambiguated corpus of 1 million words and test data is a manually created set of 958 instances. Sak et al. (2011) reports 96.45 on the same dataset of 958 manually disambiguated tokens with the use of perceptron algorithm. They also provide a morphological analyser. However, none of these studies report results on METU-Sabancı Turkish treebank data.

4 Method

The annotation of the MORPH fields in the treebank are done by applying a matching algorithm for matching the lexical forms and the tag sequences. We run the morphological analyser in two different modes as described before. Then, among the parses with tags and the lexical form output of the morphological parser, we compare the morphological tag sequence and choose the lexical form that matches the morphological tag sequence in the corresponding analysis. A lexical form may be represented with different tag sequences but this is not important since we only take the matching lexical form. We assume the morphological tag sequences are gold-standart altough as Çakıcı (2008) notes the treebank may have annotation errors in morphological disambiguation as well. For instance the first word of the example sentence in Figure 1 has a different morphological analysis assigned to it in the original treebank annotation which is corrected here. The words that could not be parsed were annotated by hand. However, the data that is created automatically by the matching algorithm need to be checked for errors caused by IG errors possibly inherent in the treebank.

Lemma field in the treebank is annotated with the stems extracted from the IGs (morphological tag sequence) for the words except verbs. The lemma for verbs are created by attaching to the extracted stem the infinitive marker *-mek* or *-mak*. The choice of the allomorph is determined by the last vowel of the extracted stem because of the vowel harmony rule in Turkish.

5 Conclusion and Future Work

In this study, we provide a treebank with complete morphological annotation. This information can be used to train systems for accurate and easier POS tagging. This can be done by various methods. One is to use a stemmer which is much more abundant in variety than morphological analysers and match the segmented data to the tags. This requires a lot less data and effort than training POS taggers that can assign the more complicated tags of the treebank directly. The use of lexical forms instead of different allomorphs or surface representation allows generalization and will prevent the sparse data problem when training these POS taggers to an extent.

None of the studies in Section 3 have reported on Turkish dependency treebank data. We aim to train automatic part of speech taggers using the segmentation data and the mapping of this segmentation to the tags in IGs using the new annotations introduced in this paper.

References

Ewan L. Antworth. 1990. *PC-KIMMO: A two-level Processor for Morphological Analysis*. Summer Institure of Linguistics, Dallas.

Nart B. Atalay, Kemal Oflazer, and Bilge Say. 2003. The annotation process in the Turkish Treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, Budapest, Hungary.

Ruket Çakıcı. 2008. *Wide-Coverage Parsing for Turkish*. Ph.D. thesis, University of Edinburgh.

Dilek Hakkani-Tür, Kemal Oflazer, and Gökhan Tür. 2002. Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities*, 36(4):381–410.

Lauri Karttunen and Kenneth R. Beesley. 2003. *Finite-State Morphology: Xerox Tools And Techniques*. CSLI Publications. Stanford University.

Kemal Oflazer and Gökhan Tür. 1996. Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 69–81.

Kemal Oflazer and Gökhan Tür. 1997. Morphological disambiguation by voting constraints. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 222–229.

Kemal Oflazer, Elvan Göçmen, and Cem Bozşahin. 1994. An outline of Turkish morphology. Technical Report TU-LANGUAGE, NATO Science Division SfS III, Brussels.

Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In Anne Abeillé and Nancy Ide, editors, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 261–277. Springer Netherlands.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 6(2).

Hasim Sak, Tunga Güngör, and Murat Saraclar. 2011. Resources for turkish morphological processing. *Language Resources and Evaluation*, 45(2):249–261.

Deniz Yüret and Ferhan Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of the Human Language Technology Conference (HLT-NAACL'06)*, pages 328–334, New York City, USA, June. Association for Computational Linguistics.