Dependency Parsing of Turkish

Gülşen Eryiğit*
Istanbul Technical University

Joakim Nivre** †
Växjö University, Uppsala University

Kemal Oflazer‡
Sabancı University

The suitability of different parsing methods for different languages is an important topic in syntactic parsing. Especially lesser-studied languages, typologically different from the languages for which methods have originally been developed, poses interesting challenges in this respect. This article presents an investigation of data-driven dependency parsing of Turkish, an agglutinative free constituent order language that can be seen as the representative of a wider class of languages of similar type. Our investigations show that morphological structure plays an essential role in finding syntactic relations in such a language. In particular, we show that employing sublexical representations called inflectional groups, rather than word forms, as the basic parsing units improves parsing accuracy. We compare two different parsing methods, one based on a probabilistic model with beam search, the other based on discriminative classifiers and a deterministic parsing strategy, and show that the usefulness of sublexical units holds regardless of parsing method. We examine the impact of morphological and lexical information in detail and show that, properly used, this kind of information can improve parsing accuracy substantially. Applying the techniques presented in this article, we achieve the highest reported accuracy for parsing the Turkish Treebank.

1. Introduction

Robust syntactic parsing of natural language is an area where we have seen a tremendous development during the last ten to fifteen years, mainly on the basis of data-driven methods but sometimes in combination with grammar-based approaches. Despite this, most of the approaches in this field have only been tested on a relatively small set of languages, mostly English but to some extent also languages like Chinese, Czech, Japanese and German.

An important issue in this context is to what extent our models and algorithms are tailored to properties of specific languages or language groups. This issue is especially pertinent for data-driven approaches, where one of the claimed advantages is portability to new languages. The results so far mainly come from studies where a

* Department of Computer Engineering, Istanbul Technical University, 34469 Istanbul, Turkey. E-mail: gulsen.cebiroglu@itu.edu.tr
** School of Mathematics and Systems Engineering, Växjö University, 35260 Växjö, Sweden. E-mail: joakim.nivre@msi.vxu.se
† Department of Linguistics and Philology, Uppsala University, Box 635, 75126 Uppsala, Sweden
‡ Faculty of Engineering and Natural Sciences, Sabancı University, 34956 Istanbul, Turkey. E-mail: oflazer@sabanciuniv.edu

Submission received: ... 2006; Accepted for publication: ... 2006

parser originally developed for English, such as the Collins parser (Collins 1997, 1999), is applied to a new language, which often leads to a significant decrease in the measured accuracy (Collins et al. 1999; Bikel and Chiang 2000; Dubey and Keller 2003; Levy and Manning 2003; Corazza et al. 2004). However, it is often quite difficult to tease apart the influence of different features of the parsing methodology in the observed degradation of performance.

A related issue concerns the suitability of different kinds of syntactic representation for different types of languages. Whereas most of the work on English has been based on constituency-based representations, partly influenced by the availability of data resources such as the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993), it has been argued that free-word order languages can be analyzed more adequately using dependency-based representations, which is also the kind of annotation found, e.g., in the Prague Dependency Treebank of Czech (Hajič et al. 2001). Recently, dependency-based parsing has been applied to a dozen different languages in the shared task of the 2006 Conference on Computational Natural Language Learning (CoNLL) (Buchholz and Marsi 2006).

In this article, we focus on dependency-based parsing of Turkish, a language that is characterized by a rich agglutinative morphology, free constituent order, and predominantly head-final syntactic constructions. Thus, Turkish can be viewed as the representative of a class of languages that are very different from English and most other languages that have been studied in the parsing literature. Using data from the recently released Turkish Treebank (Oflazer et al. 2003), we investigate the impact of different design choices in developing data-driven parsers. There are essentially three sets of issues that are addressed in these experiments.

- The first set concerns the basic *parsing methodology*, including both parsing algorithms and learning algorithms, where we contrast a statistical parser using a conditional probabilistic model with a deterministic classifier-based parser using discriminative learning.
- The second set includes issues relating to the treatment of *morphology* in syntactic parsing, which becomes crucial when dealing with languages where the most important clues to syntactic functions are often found in the morphology rather than in word order patterns. Thus, for Turkish, it has previously been shown that parsing accuracy can be improved by taking morphologically defined units rather than word forms as the basic units of syntactic structure (Eryiğit and Oflazer 2006). In this article, we corroborate these claims by showing that they hold regardless of which of the two parsers we use, and we also study the impact of different morphological feature representations on parsing accuracy.
- The third set of issues concerns *lexicalization*, a topic that has been very prominent in the parsing literature lately. Whereas the best performing parsers for English all make use of lexical information, the real benefits of lexicalization for English as well as other languages remains controversial (Klein and Manning 2003; Dubey and Keller 2003; Arun and Keller 2005).

In addition, we investigate learning curves and provide an error analysis for the best performing parser. Finally, we examine the impact of using automatically assigned part-of-speech tags instead of the manually disambiguated tags that are used in most of the experiments.

The rest of the article is structured as follows. Section 2 gives a very brief introduction to Turkish morphology and syntax and discusses the representation of morphological information and syntactic dependency relations in the Turkish Treebank. Section 3 is devoted to methodological issues, in particular the data sets and evaluation metrics used in experiments. The following two sections present two different dependency parsers trained and evaluated on the Turkish Treebank: a probabilistic parser (section 4) and a classifier-based parser (section 5). Section 6 investigates the impact of lexicalization and morphological information on the two parsers, and section 7 examines their learning curves. Section 8 presents an error analysis for the best performing parser, and section 9 analyzes the degradation in parsing performance when using automatically assigned part-of-speech tags. Section 10 discusses related work, and section 11 summarizes the main conclusions from our study.

2. Turkish: Morphology and Dependency Relations

Turkish displays rather different characteristics compared to the more well-studied languages in the parsing literature. Most of these characteristics are also found in many agglutinative languages such as Basque, Estonian, Finnish, Hungarian, Japanese and Korean. Turkish is a flexible constituent order language. Even though in written texts, the constituent order predominantly conforms to the SOV order, constituents may freely change their position depending on the requirements of the discourse context. From a dependency structure point of view, Turkish is predominantly (but not exclusively) head final.

Turkish has a very rich agglutinative morphological structure. Nouns can give rise to about one hundred inflected forms and verbs to many more. Furthermore, Turkish words may be formed through very productive derivations, increasing substantially the number of possible word forms that can be generated from a root word. It is not uncommon to find up to four or five derivations in a single word. Previous work on Turkish (Hakkani-Tür, Oflazer, and Tür 2002; Oflazer et al. 2003; Oflazer 2003; Eryiğit and Oflazer 2006) has represented the morphological structure of Turkish words by splitting them into inflectional groups (IGs). The root and derivational elements of a word are represented by different IGs, separated from each other by derivational boundaries (DB). Each IG is then annotated with its own part-of-speech and any inflectional features as illustrated in the following example:[1]

arabanızdaydı
(it was in your car)

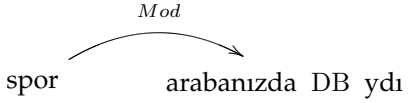
arabanızda	DB	ydı
araba+Noun+A3sg+P2pl+Loc	DB	+Verb+Zero+Past+A3sg
IG_1		IG_2
in your car		*it was*

In this example, the root of the word *arabanızdaydı* is *araba (car)* and its part-of-speech is noun. From this, a verb is derived in a separate IG. So, the word is composed of two IGs where the first one "arabanızda"(*in your car*) is a noun in locative case and in second plural possessive form and the second one is a verbal derivation from this noun, which is in past tense and third person singular form.

1 +A3sg = 3sg number agreement, +P2pl = 2pl possessive agreement, +Loc = Locative Case.

2.1 Dependency Relations in Turkish

Since most syntactic information is mediated by morphology, it is not sufficient for the parser to only find dependency relations between orthographic words; the correct IGs involved in the relations should also be identified. We can motivate this with the following very simple example: In the phrase "spor arabanızdaydı" (*it was in your sports car*), the adjective "spor" (*sports*) should be connected to the first IG of the second word. It is the word "araba" (*car*) which is modified by the adjective, not the derived verb form "arabanızdaydı" (*it was in your car*). So a parser should not just say that the first word is a dependent of the second but also state that the syntactic relation is between the last IG of the first word and the first IG of the second word as shown below.



In Figure 1 we see a complete dependency tree for a Turkish sentence laid on top of the words segmented along IG boundaries. The rounded rectangles show the words while IGs within words are marked with dashed rounded rectangles. The first thing to note in this figure is that the dependency links always emanate from the last IG of a word, since that IG determines its role as a dependent. The dependency links land on one of the IGs of a (head) word (almost always to the right). One can also note that no dependency links emanate from IGs which are not word final (e.g., the first IG of the word *okuldaki* in Figure 1). Such IGs may only have incoming dependency links and are assumed to be morphologically linked to the next IG to the right (but we do not explicitly show these links).[2]

The noun phrase formed by the three words *öğrencilerin en akıllısı* in this example highlights the importance of the IG-based representation of syntactic relations. Here in the word *akıllısı*, we have three IGs: the first contains the singular noun *akıl* (intelligence), the second IG indicates the derivation into an adjective *akıllı* (intelligence-with $\rightarrow$ intelligent). The preceding word *en* (most), an intensifier adverb, is linked to this IG as a modifier. The third IG indicates another derivation into a noun (a singular entity that is most intelligent). This last IG is the head of a dependency link emanating from the word *öğrencilerin* with genitive case-marking (of the students or students') which acts as the possessor of the last noun IG of the third word *akıllısı*. Finally, this word is the subject of the verb IG of the last word, through its last IG.

2.2 The Turkish Treebank

We have used the Turkish Treebank (Oflazer et al. 2003), created by the Middle East Technical University and Sabancı University in the experiments we report in this article. This treebank comprises 5635 sentences in which words are represented with IG-based gold-standard morphological representation and dependency links between IGs. The average number of IGs per word is 1.26 in running text, but the figure is higher for open

2 It is worth pointing out that arrows in this representations point from dependents to heads, since representations with arrows in the opposite direction also exist in the literature.

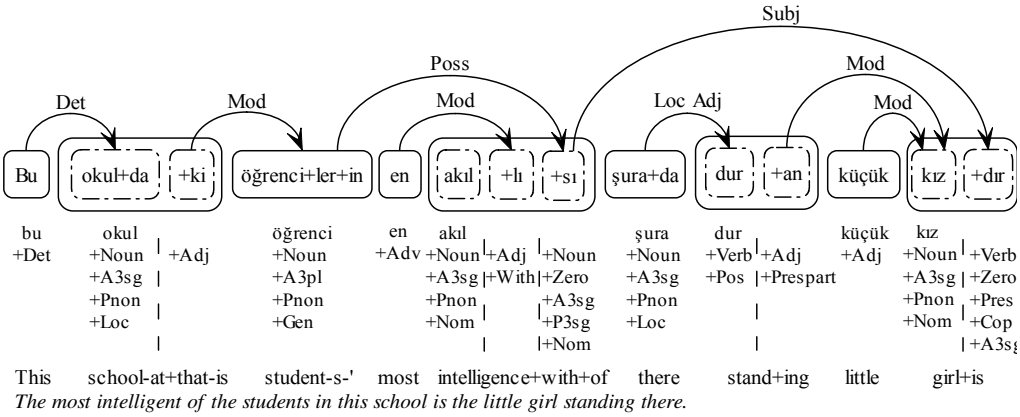


Figure 1
Dependency links in an example Turkish sentence.
+'s indicate morpheme boundaries. The rounded rectangles show the words while the inflectional groups within the words that have more than 1 IG are emphasized with the dashed rounded rectangles. The inflectional features of each inflectional group as produced by the morphological analyzer are listed below.

class words and 1 for high frequency function words which do not inflect. Of all the dependencies in the treebank, 95% are head-final and 97.5% are projective.[3]

Even though the number of sentences in the Turkish Treebank is in the same range as for many other available treebanks for languages such as Danish (Kromann 2003), Swedish (Nilsson, Hall, and Nivre 2005) and Bulgarian (Simov, Popova, and Osenova 2002), the number of words is considerably smaller (54k as opposed to 70–100k for the other treebanks). This corresponds to a relatively short average sentence length in the treebank of about 8.6 words, which is mainly due to the richness of the morphological structure, since often a word in Turkish may correspond to a whole sentence in another language.

3. Dependency Parsing of Turkish

In the upcoming sections, we investigate different approaches to dependency parsing of Turkish and show that using parsing units smaller than words improves the parsing accuracy. Below we start by describing our evaluation metrics and the data sets used, and we continue by presenting our baseline parsers: two näive parsers, which link a dependent to an IG in the next word, and one rule-based parser. We then present our data-driven parsers in the following sections: a statistical parser using a conditional probabilistic model (from now on referred to as the *probabilistic parser*) in section 4 and a deterministic classifier-based parser using discriminative learning (from now on referred to as the *classifier-based parser*) in section 5.

3 A dependency between a dependent i and a head j is projective if and only if all the words or IGs that occur between i and j in the linear order of the sentence are dominated by j. A dependency analysis with only projective dependencies corresponds to a constituent analysis with only continuous constituents.

3.1 Data Sets and Evaluation Metrics

Our initial exploration was carried out on a subset of the treebank sentences containing only projective head-final dependencies, due to the limited size of the treebank and the very small proportion of head-initial and non-projective dependencies in the treebank. In the following sections, we will evaluate our parsers both on this restricted subset and on the entire treebank including also sentences that contain head-initial or non-projective dependencies. We will refer to the two sets in tables as "*Projective HF Treebank*" and "*Entire Treebank*". The training data for parsers will never include sentences with non-projective dependencies but sometimes sentences with head-initial dependencies. This will be described separately for each of the parsers in sections 4 and 5.

We use ten-fold cross-validation for the evaluation of the parsers, except for the baseline parsers which do not need to be trained. We randomly divide the data set into ten equal parts and in each iteration use nine parts as training data and test the parser on the remaining part.

We report the results as mean scores of the ten-fold cross-validation, with standard error. The evaluation metrics used are the unlabeled attachment score (AS_U) and labeled attachment score (AS_L), i.e., the proportion of tokens that are attached to the correct head (with the correct label for AS_L). A correct attachment is one in which the dependent IG (the last IG in the dependent word) is not only attached to the correct head word *but also to the correct IG within the head word*. Where relevant, we also report the (unlabeled) word-to-word score (WW_U), which only measures whether a dependent word is connected to (some IG in) the correct head word. Dependency links emanating from punctuation are excluded in all evaluation scores. Non-final IGs of a word are assumed to link to the next IG within the word, but these links, referred to as *InnerWord* links, are not considered as dependency relations and are excluded in evaluation scoring.

3.2 Baseline Parsers

We implemented three baseline parsers to assess the performance of our probabilistic and classifier-based parsers. The first baseline parser attaches each word (from the last IG) to the first IG of the next word while the second parser attaches each word to the final IG of the next word. Obviously these two baseline parsers behave the same when the head word has only one IG. The final punctuation of each sentence is assumed to be the root of the sentence and it is not connected to any head. The first two lines of Table 1 give the unlabeled attachment scores of these parsers both on the *Projective HF Treebank* and the *Entire Treebank*. We observe that attaching the link to the first IG instead of the last one gives better results.

Table 1
Results of the baseline parsers

Parsing Model	AS_U	
	Projective HF Treebank	*Entire Treebank*
Attach-to-next (first IG)	63.9	56.0
Attach-to-next (last IG)	62.1	54.1
Rule-based	73.4	70.5

The third baseline parser is a rule-based parser that uses a modified version of the deterministic parsing algorithm by Nivre (2006). This parsing algorithm, which will be explained in detail in section 5, is a linear-time algorithm that derives a dependency graph in one left-to-right pass over the input, using a stack to store partially processed tokens and a list to store remaining input tokens in a way similar to a shift-reduce parser. In the rule-based baseline parser, the next parsing action is determined according to 31 predefined hand-written rules. The rules determine whether to connect the units (words or IGs) on top of the stack and at the head of the input list or not (regardless of dependency labels). It can be seen that the rule-based parser provides an improvement of about 10 percentage points on *Projective HF Treebank* compared to the relatively naive simpler baseline parsers. The improvement is even higher (about 15 percentage points) on *Entire Treebank*, which includes head-initial dependencies that cannot be recovered by the simpler baseline parsers.

4. Probabilistic Dependency Parser

A well-studied approach to dependency parsing is a statistical approach where the parser takes a morphologically tagged and disambiguated sentence as input, and outputs the most probable dependency tree by using probabilities induced from the training data. Such an approach comprises three components:

1. A parsing algorithm for building the dependency analyses (Eisner 1996; Sekine, Uchimoto, and Isahara 2000)
2. A conditional probability model to score the analyses (Collins 1996)
3. Maximum likelihood estimation to make inferences about the underlying probability models (Collins 1996; Chung and Rim 2004)

4.1 Methodology

The aim of the probabilistic model is to assign a probability to each candidate dependency link by using the frequencies of similar dependencies computed from a training set. The aim of the parsing algorithm is then to explore the search space in order to find the most probable dependency tree. This can be formulated with Equation 1 where S is a sequence of n units (words or IGs) and T ranges over possible dependency trees consisting of dependency links $dep(u_i, u_{H(i)})$, with $u_{H(i)}$ denoting the head unit to which the dependent unit u_i is linked and the probability of a given tree is the product of the dependency links that it comprises.

$$T^* = \underset{T}{\text{argmax}}\ P(T|S) = \underset{T}{\text{argmax}} \prod_{i=1}^{n-1} P(dep\,(u_i, u_{\mathcal{H}(i)})\,|\,S) \tag{1}$$

The observation that 95% of the dependencies in the Turkish treebank are head-final dependencies motivated us to employ the backward beam search dependency parsing algorithm by Sekine, Uchimoto, and Isahara (2000), but adapted to our morphological representation with IGs. This algorithm is originally designed for Japanese, another head-final language. It parses a sentence starting from the end moving towards the beginning, trying at each step to link the dependents to a unit to the right. It uses a beam which keeps track of the most probable dependency structures for the partially

processed sentence. Since this algorithm makes the projectivity assumption, we limited ourself to only work on sentences with projective head-final dependencies, i.e., on *Projective HF Treebank*. This means that both training and test data is restricted to sentences containing only projective head-final dependencies.

For the probability model, we adopt the approach by Chung and Rim (2004), which itself is a modified version of the statistical model used in Collins (1996).[4] In this model in Equation 2, the probability of a dependency link $P(dep\,(u_i, u_{\mathcal{H}(i)}) \,|\, S)$ from u_i linking to a head $u_{H(i)}$, is approximated with the product of two probabilities:

$$P(dep\,(u_i, u_{\mathcal{H}(i)}) \,|\, S) \approx P(link(u_i, u_{\mathcal{H}(i)}) \,|\, \Phi_i\ \Phi_{\mathcal{H}(i)}) \cdot \\ P(u_i \text{ links to some head } dist(i, H(i)) \text{ away} \,|\, \Phi_i) \tag{2}$$

In this equation

- $P(link(u_i, u_{\mathcal{H}(i)}) \,|\, \Phi_i\ \Phi_{\mathcal{H}(i)})$ is the probability of seeing the same dependency within a similar context where Φ_i represents the context around the dependent u_i and $\Phi_{\mathcal{H}(i)}$ represents the context around the head $u_{H(i)}$, and
- $P(u_i \text{ links to some head } dist(i, H(i)) \text{ away} \,|\, \Phi_i)$ is the probability of seeing the dependent linking to some head a distance $dist(i, H(i))$ away, in the context Φ_i.

In all of the following models, $dist(i, H(i))$ is taken as the number of actual word boundaries between the dependent and the head unit regardless of whether full words or IGs were used as units of parsing.[5]

For smoothing the probabilities in Equation 2, we used a modified version of the backed-off smoothing used by Collins (1996). We interpolated the probabilities in equation 2 calculated from the treebank by removing the head and the dependent contextual information all at once.[6] So, during the actual runs the smoothed probability $P(link(u_i, u_{\mathcal{H}(i)}) \,|\, \Phi_i\ \Phi_{\mathcal{H}(i)})$ is computed by interpolating two unsmoothed estimates extracted from the treebank: $P_1(link(u_i, u_{\mathcal{H}(i)}) \,|\, \Phi_i\ \Phi_{\mathcal{H}(i)})$ and $P_2(link(u_i, u_{\mathcal{H}(i)}))$. A similar approach was employed for $P(u_i \text{ links to some head } dist(i, H(i)) \text{ away} \,|\, \Phi_i)$. If even after interpolation, the probability is 0, then a very small value is used. Further distances larger than a certain threshold value were assigned the same probability, as explained later.

4.2 The Choice of Parsing Units

In the probabilistic dependency parsing experiments, we experimented with three different ways of choosing and representing the units for parsing:

4 The statistical model in Collins (1996) is actually used in a phrase-structure-based parsing approach, but it uses the same idea of computing probabilities between dependents and head units. We also tried to employ this statistical model by Collins where the distance measure is included in the probability formula, but we obtained worse results with this.

5 We also tried other distance functions, e.g., the number of IGs between dependent and head units, but this choice fared better than the alternatives.

6 We tried many other backed-off models such as removing the neighbors one by one or removing the inflectional features. But we obtained the best results by removing all the neighbors together.

1. **Word-based Model #1**: In this model, the units of parsing are the actual words and each word is represented by a combination of the representations of *all* the IGs that make it up. Note that although all IGs are used in representing a word, not all the information provided by an IG has to be used as we will see shortly. This representation however raises the following question: If we use the words as the parsing units and find the dependencies between these, how can we translate these to the dependencies between the IGs, since our goal is to find dependencies between IGs? The selection of the IG of the dependent word is an easy decision, as it is the last IG in the word. The selection of the head IG is obviously more difficult. Since such a word-based model will not provide much information about the underlying IGs structure, we will have to make some assumptions about the head IG. The observation that 85.6% of the dependency links in the treebank land on the first (and possibly the only) IG of the head word and the fact that our first baseline model (attaching to the first IG) gives better performance than our second baseline model (attaching to the last IG), suggest that after identifying the correct word choosing the first IG as the head IG may be a reasonable heuristic. Another approach to determining the correct IG in the head word could be to develop a post-processor which selects this IG using additional rules. Such a post-processor could be worth developing if the WW_U accuracy obtained with this model proves to be higher than all of the other models, i.e., if this is the best way of finding the correct dependencies between words without considering which IGs are connected. However, as we will see in section 4.4, this model does not give the best WW_U.
2. **Word-based model #2**: This model is just like the model above but we represent a word using its final IGs rather than the concatenation of all their IGs when *it is used as a dependent*. The representation is the same as in Word-based model #1 when the word is a head. This results in a dynamic selection of the representation during parsing as the representation of a word will be determined according to its role at that moment. The representation of the neighboring units in context will again be selected with respect to the word in question: any context unit on the left will be represented with its dependent representation (just the last IG) and any neighbor on the right will be represented with its representation as a head. The selection of the IG in the head word is the same as in the first model above.
3. **IG-based model**: In this model, we use IGs as units in parsing. So we split the IG-based representation of each word and reindex these IGs in order to use them as single units in parsing. Figure 2 shows this transfer to the IG-based model. We still however need to know which IGs are word-final as they will be the dependent IGs (shown in the figure with * superscript). The contextual elements that are used in this model are the IGs to the left (starting with the last IG of the preceding word) and the right of the dependent and the head IG.

4.3 Reduced Dynamic Representations for IGs

In all the models above, it is certainly possible to use all the information supplied by the full morphological analysis in representing the IGs.[7] This includes the root words themselves, major and minor[8] parts of-speech, number and person agreement markers, possessive agreement markers, case markers, tense, aspect, mood marker and other

7 See Figure 1 for a sample for such information.
8 A minor part-of-speech category is available for some major part-of-speech categories: e.g., pronouns are further divided into personal pronouns, demonstrative pronouns, interrogative pronouns, etc. The minor part-of-speech category always implies the major part-of-speech. For derived IGs the minor

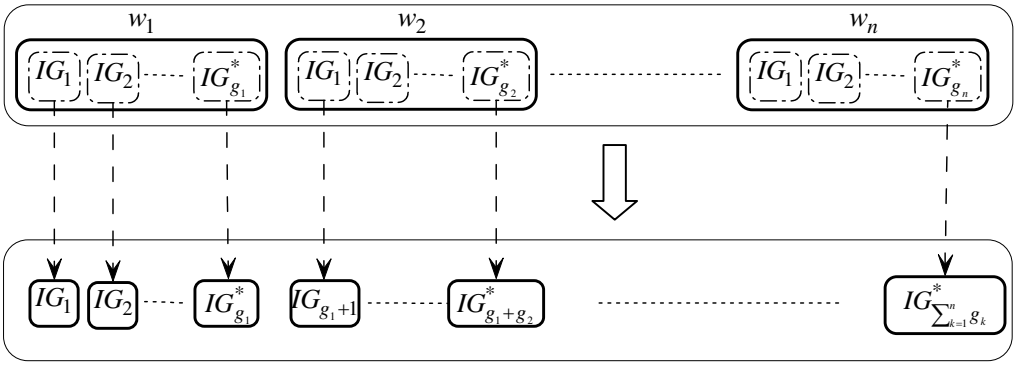


Figure 2
Conversion to the IG-based representation

miscellaneous inflectional and semantic markers especially for derivations. Not all of these features may be relevant to the parsing task, and further, different features may be relevant depending on whether the IG is being used as a dependent or a head. Also in order to alleviate the data sparseness problem that may result from the relatively modest size of the treebank, an "unlexicalized" representation that does not contain the root word needs to be considered so that statistics from IGs that are otherwise same except for the root word (if any) can be conflated.[9]

After some preliminary experimentation, we decided that a reduced representation for IGs that is dynamically selected depending on head or dependent status would give us the best performance.

Below, we explain the representation of the IGs and the parameters that we used in the three models above.

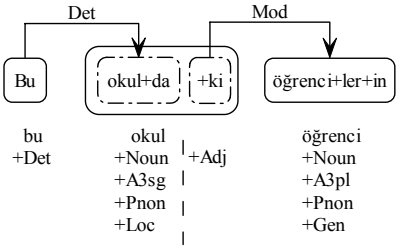
- When used as a dependent (or part of a dependent word in models 1 and 2) during parsing
  - Nominal IGs (nouns, pronouns, and other derived forms that inflect with the same paradigm as nouns, including infinitives, past and future participles) are represented only with the case marker, since that essentially determines the syntactic function of that IG as a dependent, and only nominals have cases.
  - Any other IG is just represented with its minor part-of-speech.
- When used as a head (or part of a head word in models 1 and 2)
  - Nominal IGs and adjective IGs with participle minor part-of-speech[10] are represented with the minor part-of-speech and the possessive agreement marker.
  - Any other IG is just represented with its minor part-of-speech.

part-of-speech mostly indicates a finer syntactic or semantic characterization of the derived word. When no minor part-of-speech is available the major part-of-speech is used.

9 Remember that only the first IG in a word has the root word.

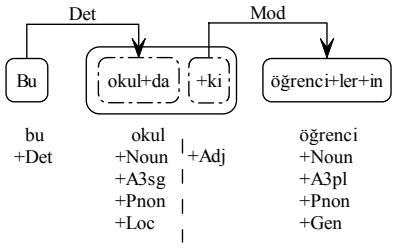
10 These are modifiers derived from verbs. They have adjective as their major part-of-speech and past/future participle as their minor part-of-speech. They are the only types of IGs that have possessive agreement marker other than nominals.

Figures 3 – 5 shows for the first three words in Figure 1, the unlexicalized reduced representations that are used in the three models above, when units are used as dependents and heads during parsing.



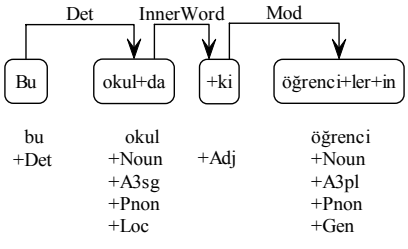
dependent	<+Det>	<+Loc+Adj>	<+Gen>
head	<+Det>	<+Noun+Pnon+Adj>	<+Noun+Pnon>

Figure 3
Reduced IG representation for Word-based model #1



dependent	<+Det>	<+Adj>	<+Gen>
head	<+Det>	<+Noun+Pnon+Adj>	<+Noun+Pnon>

Figure 4
Reduced IG representation for Word-based model #2



dependent	<+Det>	<+Loc>	<+Adj>	<+Gen>
head	<+Det>	<+Noun+Pnon>	<+Adj>	<+Noun+Pnon>

Figure 5
Reduced IG representation for IG-based model

4.4 Experimental Results

In this section, we first evaluate the performance of the models described in section 4.2. We then investigate the impact of different choices on the IG-based model.

In addition to the parsing model, the parser is given the following parameters:

- the number of left and right neighbors of the dependent (D_l, D_r) to define the dependent context Φ_i,
- the number of left and right neighbors of the head (H_l, H_r) to define the head context $\Phi_{H(i)}$,
- the size of the beam ($beamsize$),
- the distance *threshold* value beyond which $P(u_i \text{ links to some head } dist(i, H(i)) \text{ away} \mid \Phi_i)$ is assigned the same probability.

Table 2 gives the AS_U scores for the word-based and IG-based models together with the optimized parameters. For all three models, the *beamsize* value is selected as 3 and *threshold*[11] is selected as 6 which are observed to give the best performance. It can be seen that the performance of the word-based models is lower than our rule-based baseline parser (Table 1) with AS_U=73.4, even though they are better than the first two rather naive baselines. On the other hand, the IG-based model outperforms all of the baseline parsers and word-based models. It should also be noted that the IG-based model not only improves the AS_U accuracy but also the word-to-word accuracy. Thus, the IG-based model not only helps to recover the relations between correct IGs but also to find the correct head word.

Table 2
Results of the probabilistic parser

Parsing Model *(parameters)*	AS_U (*Projective HF Treebank*)	WW_U (*Projective HF Treebank*)
Word-based model #1 *(D_l=1, D_r=1, H_l=1, H_r=1)*	71.5$\pm$0.5	78.5$\pm$1.3
Word-based model #2 *(D_l=1, D_r=1, H_l=1, H_r=1)*	72.0$\pm$0.4	79.1$\pm$1.1
IG-based model *(D_l=1, D_r=1, H_l=0, H_r=1)*	74.9$\pm$0.3	82.2$\pm$0.8

In Table 3, we also presents results from experiments employing different representations for the IGs. A more detailed investigation about using limited lexicalization and inflectional features will be presented later in section 6. Here, we will see what would have happened if we had used alternative reduced IG representations compared to the representation described earlier, which is used in the best performing IG-based model.

Table 3 gives the results for each change to the representational model. One can see that none of these representational changes improves the performance of the best

11 As stated earlier, our distance function is calculated according to the word boundaries between the dependent and the head units. In the treebank, 95% of the dependency links link to a word that is less than 6 words away. Thus all the distances larger than or equal to 6 are conflated into the same small probability.

Table 3
Results for different representations

#	Model (D_l=1, D_r=1, H_l=0, H_r=1) (*Projective HF Treebank*)	AS_U
	IG-based model	74.9±0.3
1	Using major part-of-speech instead of minor part-of-speech	74.6±0.3
2	Using only minor part-of-speech and no other inflectional features	72.2±0.4
3	Using minor part-of-speech for all types of IGs together with case and possessive markers for nominals and possessive marker for adjectives (but no dynamic selection)	73.0±0.4
4	Using all inflectional features in addition to minor part-of-speech	50.4±0.5
5	Adding the root information to best performing IG-based model	57.8±0.4
6	Adding surface form information to best performing IG-based model	62.3±0.4

performing model. Only employing major part-of-speech tags (#1) actually comes close, and the difference is not statistically significant. Lexicalization of the model results in a drastic decrease in performance: using the surface form (#6) gives somewhat better results than using root information (#5). Also dynamic selection of tags seems to help performance (#3) but using all available inflectional information performs significantly worse possibly due to data sparseness.

5. Classifier-based Dependency Parser

Our second data-driven parser is based on a parsing strategy whose success is reported to be very high across a variety of different languages (Nivre et al. 2006). This strategy consists of the combination of the following three techniques:

1. Deterministic parsing algorithms for building dependency graphs (Kudo and Matsumoto 2002; Yamada and Matsumoto 2003; Nivre 2003)
2. History-based models for predicting the next parser action (Black et al. 1992; Magerman 1995; Ratnaparkhi 1997; Collins 1999)

3. Discriminative classifiers to map histories to parser actions (Veenstra and Daelemans 2000; Kudo and Matsumoto 2002; Yamada and Matsumoto 2003; Nivre, Hall, and Nilsson 2004)

A system of this kind employs no grammar but relies completely on inductive learning from treebank data for the analysis of new sentences, and on deterministic parsing for disambiguation. This combination of methods guarantees that the parser is robust, never failing to produce an analysis for an input sentence, and efficient, typically deriving this analysis in time that is linear or quadratic in the length of the sentence.

In the following subsections, we will first present the parsing methodology, and then we will present results that show that the IG-based model again outperforms the word-based model. We will then explore how we further improve the success by exploiting the advantages of this parser.

5.1 Methodology

For the experiments in this article, we use a variant of the parsing algorithm proposed by Nivre (2003, 2006), a linear-time algorithm that derives a labeled dependency graph in one left-to-right pass over the input, using a stack to store partially processed tokens and a list to store remaining input tokens. However, in contrast to the original arc-eager parsing strategy, we use an arc-standard bottom-up algorithm, as described in Nivre (2004). Like most of the algorithms used for practical dependency parsing, this algorithm is restricted to projective dependency graphs.

The parser uses two elementary data structures, a stack σ of partially analyzed tokens and an input list τ of remaining input tokens. The parser is initialized with an empty stack and with all the tokens of a sentence in the input list; it terminates as soon as the input list is empty. In the following, we use subscripted indices, starting from 0, to refer to particular tokens in σ and τ. Thus, σ_0 is the token on top of the stack σ (the *top token*) and τ_0 is the first token in the input list τ (the *next token*); σ_0 and τ_0 are collectively referred to as the *target tokens*, since they are the tokens considered as candidates for a dependency relation by the parsing algorithm.

There are three different parsing actions, or transitions, that can be performed in any non-terminal configuration of the parser:

- **Shift**: Push the next token onto the stack.
- **Left-Arc**$_r$: Add a dependency arc from the next token to the top token, labeled r, then pop the stack.
- **Right-Arc**$_r$: Add a dependency arc from the top token to the next token, labeled r, then replace the next token by the following token at the head of the input list.

In order to perform deterministic parsing in linear time, we need to be able to predict the correct parsing action (including the choice of a dependency type r for **Left-Arc**$_r$ and **Right-Arc**$_r$) at any point during the parsing of a sentence. This is what we use a history-based classifier for.

The features of the history-based model can be defined in terms of different linguistic features of tokens, in particular the target tokens. In addition to the target tokens, features can be based on neighboring tokens, both on the stack and in the remaining

input, as well as dependents or heads of these tokens in the partially built dependency graph. The linguistic attributes available for a given token are the following:

- Lexical form (root) (LEX)
- Part-of-speech category (POS)
- Inflectional features (INF)
- Dependency type to the head if available (DEP)

To predict parser actions from histories, represented as feature vectors, we use support vector machines (SVM), which combine the maximum margin strategy introduced by Vapnik (1995) with the use of kernel functions to map the original feature space to a higher-dimensional space. This type of classifier has been used successfully in deterministic parsing by Kudo and Matsumoto (2002), Yamada and Matsumoto (2003), and Sagae and Lavie (2005), among others. To be more specific, we use the LIBSVM library for SVM learning (Chang and Lin 2001), with a polynomial kernel of degree 2, with binarization of symbolic features, and with the one-versus-one strategy for multi-class classification.[12]

This approach has some advantages over the probabilistic parser, in that

- it can process both left-to-right and right-to-left dependencies due to its parsing algorithm,
- it assigns dependency labels simultaneously with dependency and can use these as features in the history-based model,
- it does not necessarily require expert knowledge about the choice of linguistically relevant features to use in the representations since SVM training involves implicit feature selection.

To compare the results with the previous ones, we evaluate the classifier-based parser on both datasets (on *Projective HF Treebank* and on *Entire Treebank*). When testing on *Projective HF Treebank*, we use the same training data as for the probabilistic parser, that is, sentences containing only projective head-final dependencies. When testing on *Entire Treebank*, we also include in the training data sentences containing head-initial dependencies. However, we still exclude sentences with non-projective dependencies during training.[13] Since the classifier-based parser not only builds dependency structures but also assigns dependency labels, we give AS_L scores as well as AS_U scores.

5.2 Experimental Results

In this section, our first aim is to confirm the claim that using IGs as the units in parsing improves performance. For this purpose, we start by using models similar to those described in the previous section. We use an unlexicalized feature model where the parser uses only the minor part-of-speech category (POS) and the dependency types (DEP) features for the tokens and compare the results with the probabilistic parser.

12 Experiments have also been performed using memory-based learning (Daelemans and Bosch 2005). They however gave a lower parsing accuracy.

13 Since the frequency of non-projective dependencies in the Turkish Treebank is not high enough to learn such dependencies, we did not observe any improvement when applying the pseudo-projective processing of Nivre and Nilsson (2005), which is reported to improve accuracy for other languages.

We then show in the second part how we can improve accuracy by exploiting the morphological structure of Turkish and taking advantage of the special features of this parser.

5.2.1 Comparison with the Probabilistic Parser. In order to compare with the results of the previous section, we adopt the same strategy that we used earlier in order to present inflectional groups. We employ two representation models:

- **Word-based model**, where each word is represented by the concatenation of its IGs,
- **IG-based model**, where the units are inflectional groups.

We take the minor part-of-speech category plus the case and possessive agreement markers for nominals and participle adjectives to make up the POS feature[14] of each IG. However, we do not employ dynamic selection of these features and just use the same strategy for both dependents and the heads. The reason is that, in this parser, we do not make the assumption that the head is always on the right side of the dependent but also try to find head-initial dependencies and the parser does not know at a given stage if a unit is a candidate head or dependent. In the IG-based model, *InnerWord* relations (Figure 5), which are actually determined by the morphological analyzer, are processed deterministically without consulting the SVM classifiers.[15]

The feature model (Feature Model #1) to be used in these experiments is shown in Figure 6. This feature model uses five POS features, defined by the POS of the two topmost stack tokens (σ_0, σ_1), the first two tokens of the remaining input (τ_0, τ_1) and the token which comes just after the topmost stack token in the actual sentence ($\sigma_0 + 1$). The dependency type features involve the top token on the stack (σ_0), its leftmost and rightmost dependent ($l(\sigma_0)$, $r(\sigma_0)$), and the leftmost dependent of the next input token ($l(\tau_0)$).

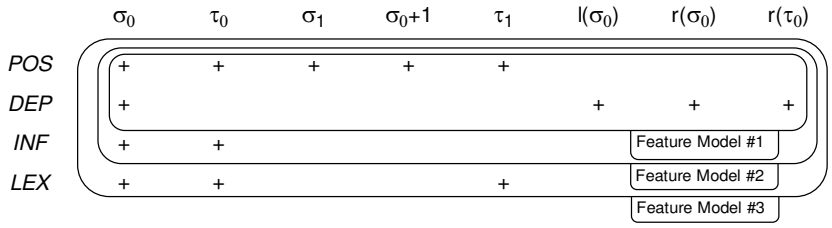


Figure 6
Feature models for the classifier-based parser

The results for this feature model and the two representation models can be seen in Table 4. We again see that the IG-based model outperforms the word-based model. When we compare the unlabeled (AS_U) scores on *Projective HF Treebank* with the results

14 Thus, we are actually combining some inflectional features with the part-of-speech category and use them together in the POS feature.

15 For the unlexicalized models, it is necessary to process *InnerWord* relations deterministically in order to get the full benefit of IG-based parsing, since the classifiers cannot correctly predict these relations without lexical information (Eryiğit, Nivre, and Oflazer 2006). However for the lexicalized models, adding deterministic *InnerWord* processing has no impact at all on parsing accuracy, but it reduces training and parsing time by reducing the number of training instances for the SVM classifiers.

of the probabilistic parser (from Table 2), we see that we do not obtain any improvements for the IG-based model and even have a significant decrease in the word-based model. This may be related to not using dynamic selection.[16]

Table 4
Results for the unlexicalized classifier-based parser

Parsing Model	*Projective HF Treebank*		*Entire Treebank*	
	AS_U	AS_L	AS_U	AS_L
Word-based model	70.5±0.5	60.7±0.5	67.1±0.3	57.8±0.3
IG-based model	74.6±0.3	64.2±0.4	70.6±0.2	60.9±0.3

5.2.2 Exploiting the Advantages of the Classifier-based Parser. To exploit the advantages of the classifier-based parser, we now describe a setting which does not rely on any linguistic knowledge on the selection of inflectional features and lets the classifier of the parser select the useful combinations of the features. As support vector machines can perform such tasks successfully, we now explore different representations of the morphological data in the IG-based model to see if the performance can be improved.

As shown in earlier examples, the inflectional information available for a given token normally consists of a complex combination of atomic features such as +A3sg, +Pnon and +Loc. Recent work (Eryiğit, Nivre, and Oflazer 2006) showed that adding inflectional features as atomic values to the feature models, was better than taking certain subsets with linguistic intuition and trying to improve on them. Thus we now present results with the feature model where the POS component only comprises the minor part-of-speech and the INF comprises all the other inflectional features provided by the treebank without any reduction. We investigate the impact of this approach first with an unlexicalized model (Feature Model #2 in Figure 6) and then with a lexicalized model (Feature Model #3 in Figure 6) where we investigate two different kinds of lexicalization: one using just the root information and one using the complete surface form as lexical features.

Table 5 gives the results for both unlexicalized and lexicalized models with INF features included in the feature model. We can see the benefit of using inflectional features separately and split into atomic components by comparing the first line of the table with the best results for the IG-based model in Table 4. We can also note the improvement that lexicalized models bring:[17] In contrast to the probabilistic parser, lexicalization using root information rather than surface form gives better performance even though the difference is not statistically significant. The improvement of AS_U scores on the (*Projective HF Treebank*) is 3.4 percentage points for the lexicalized model (with root) and 2.2 for the unlexicalized model over the IG-based model of the probabilistic parser with AS_U=74.9±0.3. Thus, the improvement in accuracy cannot be attributed to lexicalization alone. A similar case can be observed for WW_U accuracies: Including INF and lexicalization with roots gives WW_U=85.5±1.0 on *Projective HF Treebank*, which provides

16 Actually, the equivalent of this IG-based model is the probabilistic model #3 in Table 3 (with no dynamic selection) which does significantly worse than this classifier-based model.

17 The unlabeled exact match score (that is, the percentage of sentences for which *all* dependencies are correctly determined) for this best performing model on *Entire Treebank* is 37.5% upon IG-based evaluation and 46.5% upon word-based evaluation.

an improvement of 3.3 percentage points over the IG-based model of the probabilistic parser (with WW_U=82.2$^{\pm 0.8}$). The WW_U accuracy of the unlexicalized Feature Model #2 is 82.8$^{\pm 1.2}$, which is slightly better but not statistically significantly.

Table 5
Results for enhancements of the IG-based model

Feature Model	*Projective Hf Treebank*		*Entire Treebank*	
	AS_U	AS_L	AS_U	AS_L
Feat. Model #2 (no lexicalization)	76.1$^{\pm 0.3}$	65.9$^{\pm 0.4}$	72.4$^{\pm 0.2}$	63.1$^{\pm 0.3}$
Feat. Model #3 (lex. with surface forms)	78.0$^{\pm 0.4}$	68.3$^{\pm 0.3}$	75.7$^{\pm 0.2}$	66.6$^{\pm 0.3}$
Feat. Model #3 (lex. with roots)	78.3$^{\pm 0.3}$	68.9$^{\pm 0.2}$	76.0$^{\pm 0.2}$	67.0$^{\pm 0.3}$

6. The Impact of Inflectional Features and Lexicalization

In the previous sections, we have presented our parsers using optimized parameters and feature representations. We have observed that using complete inflectional features and lexicalized models improves the accuracy of the classifier-based parser significantly, whereas for the probabilistic parser adding these features has a negative impact on accuracy. In this section, we investigate the influence of different inflectional features and lexical information on both parsers using the best performing IG-based models, in order to get a more fine-grained picture. The results of the experiments with the classifier-based parser are not strictly comparable to those of other experiments, since the training data have here been divided into smaller sets (based on the major part-of-speech category of the next token) as a way of reducing SVM training times without a significant decrease in accuracy. For the probabilistic parser, we have not used dynamic selection while investigating the impact of inflectional features.

6.1 Inflectional Features

In order to see the influence of inflectional features, we tested six different sets, where each set includes the previous one and adds some more inflectional features. The following list describes each set in relation to the previous one:

Set 1 No inflectional features except for minor part-of-speech
Set 2 Set 1 + Case and possessive markers for nominals, possessive markers for participle adjectives
Set 3 Set 2 + person/number agreement features for nominals and verbs
Set 4 Set 3 + all inflectional features for nominals
Set 5 Set 4 + all inflectional features for verbs
Set 6 Set 5 + all inflectional features

Figure 7 shows the results for both the probabilistic and the classifier-based parser. The results shown in Figures 7-b and 7-c confirm the importance of case and possessive features, which was presupposed in the manual selection of features in section 4. Besides these, the number/person agreement features available for nominals and verbs are also important inflectional features even though they do not provide any statistically

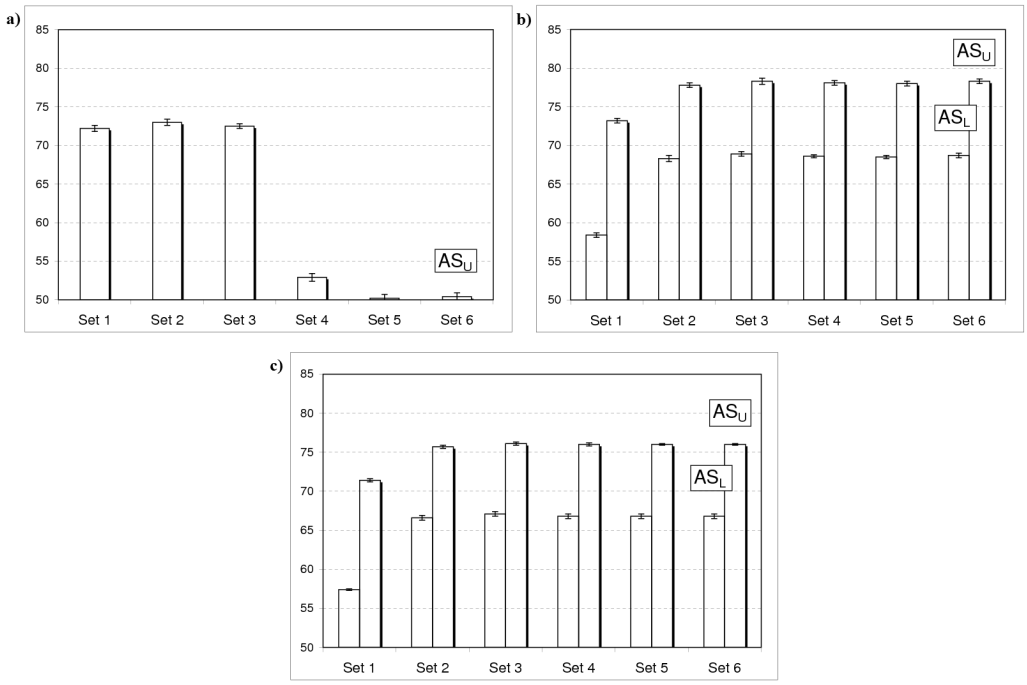


Figure 7
Accuracy for feature sets 1–6:
a) Unlabeled accuracy for probabilistic parser on *Projective HF Treebank*
b) Unlabeled and labeled accuracy for classifier-based parser on *Projective HF Treebank*
c)Unlabeled and labeled accuracy for classifier-based parser on *Entire Treebank*

significant increase in accuracy (except for AS_U in Figure 7-c (Set 3)). Another point that merits attention is the fact that the labeled accuracy is affected more by the usage of inflectional features than the unlabeled accuracy. The difference between Set 1 and Set 2 (in Figures 7-b and 7-c) is nearly 4 percentage points for AS_U and 10 percentage points for AS_L. It thus appears that the inflectional features are especially important in order to determine the type of the relationship between the dependent and head units. This is logical since in Turkish it is not the word order that determines the roles of the constituents in a sentence but the inflectional features (especially the case markers). We again see from these figures that the classifier-based parser does not suffer from sparse data even if we use the full set of inflectional features (Set 6) provided by the treebank whereas the probabilistic parser starts having this problem even with Set 3 (Figure 7-a). The problem gets worse when we add the complete inflectional features.

6.2 Lexicalization

In order to get a more fine-grained view of the role of lexicalization, we have investigated the effect of lexicalizing IGs from different major part-of-speech categories. We expand this analysis into minor part-of-speech categories where relevant. The results are shown in Table 6, where the first column gives the part-of-speech tag of the lexicalized units, while the second and third columns give the total frequency and the frequency of distinct roots for that part-of-speech tag. We again see that the probabilistic parser

suffers from sparse data especially for the tags having a high frequency of distinct roots. We cannot observe any increase with the lexicalization of any category. The situation is different for the classifier-based parser. None of the individual lexicalizations causes a decrease. We see that on *Projective HF Treebank* the lexicalization of nouns causes a significant increase in accuracy. Lexicalization of verbs also gives a noticeable increase in the labeled accuracy even though this is not statistically significant. A further investigation on the minor parts-of-speech of nouns[18] shows that only common nouns has this positive effect, whereas the lexicalization of proper nouns does not improve accuracy. On *Entire Treebank*, we see that the lexicalization of conjunctions also improves the accuracy significantly. This improvement, which is not observed on *Projective HF Treebank*, can be attributed to the enclitics (such as "de", "ki", "mi", written on the right side of and separately from the word they attach to), which give rise to head-initial dependencies that do not exist in *Projective HF Treebank*. These enclitics, which are annotated as conjunctions in the treebank, can be differentiated from other conjunctions by lexicalization which makes it very easy to connect them to their head on the left.

Table 6
Results for limited lexicalization (n = count, d = number of distinct roots)

			Probabilistic	Classifier-based			
			Projective Hf Treebank	*Projective Hf Treebank*		*Entire Treebank*	
	n	*d*	AS_U	AS_U	AS_L	AS_U	AS_L
None	-	-	*74.9*±0.3	*76.5*±0.3	*66.1*±0.3	*72.8*±0.2	*63.2*±0.3
Adjectives	6446	735	72.2±0.3	76.5±0.3	66.1±0.3	72.9±0.2	63.2±0.3
Adverbs	3033	221	74.8±0.3	76.6±0.3	66.3±0.3	73.1±0.2	63.4±0.3
Conjunctions	2200	44	70.4±0.4	76.6±0.3	66.2±0.3	**74.1**±0.2	**64.2**±0.3
Determiners	1998	13	74.8±0.3	76.6±0.3	66.1±0.3	72.8±0.2	63.3±0.3
Duplications	11	9	74.9±0.3	76.5±0.3	66.1±0.3	72.8±0.2	63.2±0.3
Interjections	100	34	74.9±0.3	76.5±0.3	66.1±0.3	72.8±0.2	63.2±0.3
Nouns	21860	3935	60.8±0.5	**77.2**±0.2	**67.1**±0.2	**73.9**±0.2	**64.6**±0.3
Numbers	850	226	70.4±0.4	76.5±0.3	66.1±0.3	72.9±0.2	63.3±0.3
Post-positions	1250	46	73.6±0.4	76.6±0.3	66.2±0.3	72.9±0.2	63.2±0.3
Pronouns	2145	28	75.0±0.3	76.5±0.3	66.1±0.3	72.8±0.2	63.2±0.3
Punctuations	10420	16	75.1±0.3	77.1±0.4	66.6±0.3	73.4±0.2	63.7±0.3
Questions	228	6	74.9±0.3	76.5±0.3	66.1±0.3	72.8±0.2	63.2±0.3
Verbs	14641	1256	67.8±0.5	76.5±0.3	**66.6**±0.2	72.9±0.2	**63.8**±0.3

Since we did not observe any improvement in the probabilistic parser, we continued further experimentation only with the classifier-based parser. We tried partially lexicalized models by lexicalizing various combinations of certain part-of-speech categories (see Figure 8). The results show that, whereas lexicalization certainly improves parsing accuracy for Turkish, only the lexicalization of conjunctions and nouns together has an

18 IGs with a noun part-of-speech tag other than common nouns are marked with an additional minor part-of-speech that indicates whether the nominal is a proper noun or a derived form – one of future participle, past participle, infinitive, or a form involving a zero-morpheme derivation. The latter four do not contain any root information.

impact on accuracy. Similar to the experiments on inflectional features, we again see that the classifier-based parser has no sparse data problem even if we use a totally lexicalized model.

Although the effect of lexicalization has been discussed in several studies recently (Klein and Manning 2003; Dubey and Keller 2003; Arun and Keller 2005), it is often investigated as an all-or-nothing affair (except for few studies which analyzes the distributions of lexical items, e.g., Bikel (2004), Gildea (2001)). The results for Turkish clearly show that the effect of lexicalization is not uniform across syntactic categories, and that a more fine-grained analysis is necessary to determine in what respects lexicalization may have a positive or negative influence. For some models (especially suffering from sparse data), it may even be a better choice to use some kind of limited lexicalization instead of full lexicalization, although the experiments in this article do not show any example of that. The results from the previous section suggests that the same is true for morphological information, but this time showing that limited addition of inflectional features (instead of using them fully) helps to improve the accuracy of the probabilistic parser.

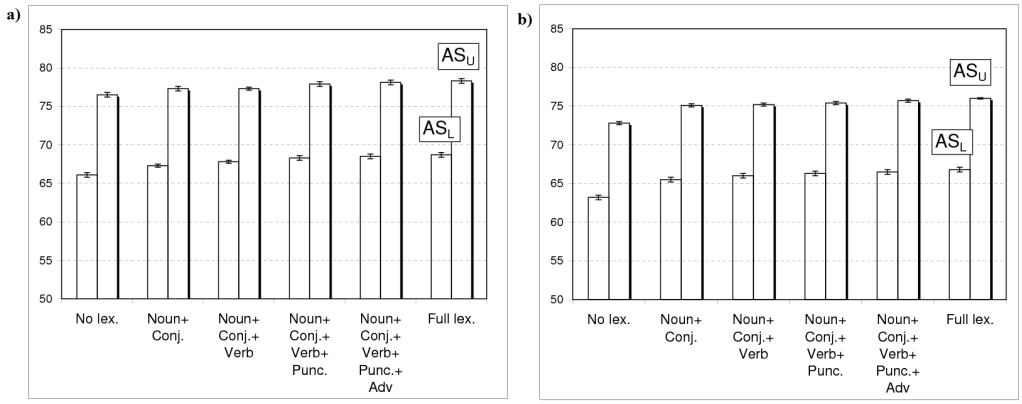


Figure 8
Results for incremental lexicalization for the classifier-based parser
a) on *Projective HF Treebank*, b) on *Entire Treebank*

7. The Impact of Training Set Size

In order to see the influence of the training set size on the performance of our parsers, we designed the experiments shown in Figure 9. In order to use the exact same training sets, we evaluated our parsers with AS_U scores on *Projective HF Treebank*. Figure 9 gives the accuracies for the probabilistic parser (unlexicalized) and the classifier-based parser (unlexicalized and lexicalized). The x-axis shows the number of cross validation subsets that we used for training in each step. We observe that the relative improvement with growing training set size is largest for the classifier-based lexicalized model with a relative difference of $4.8^{\pm 0.1}$ between using 9 training subsets and 1 training subset, whereas this number is $3.9^{\pm 0.2}$ for the unlexicalized classifier-based model and $2.7^{\pm 0.1}$ for the unlexicalized probabilistic model. We can state that despite its lower accuracy, the probabilistic model is less affected by the size of the training data. But for any of the sizes, the relative ranking of the models remain the same, except for size 1, where there is no significant difference between the performances of the unlexicalized probabilistic

and classifier-based models. Another conclusion may be that classifier-based models are better at extracting information with the increasing size of the data in hand, whereas the probabilistic model cannot be improved very much with increasing size of the data. We can observe this situation especially in the lexicalized model which is improved significantly between size=6 subsets and size=9 subsets, whereas there is no significant improvement on the unlexicalized models within this interval.

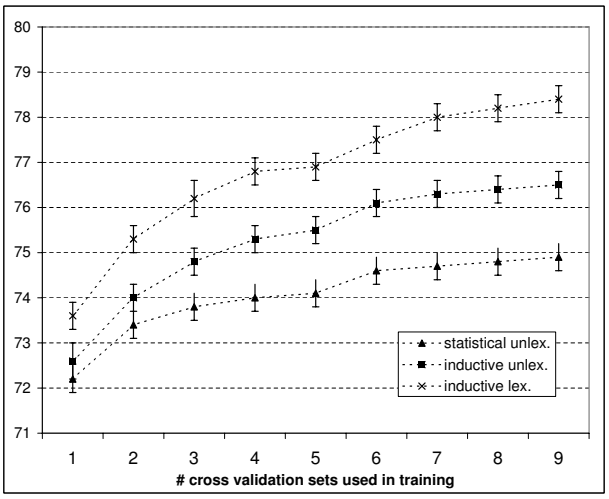


Figure 9
Unlabeled accuracy on *Projective HF Treebank* for different training set sizes

8. Error Analysis

In this section, we present a detailed error analysis on the results of our best performing parser on *Entire Treebank*. We first evaluate our results on different dependency types. We then investigate the error distribution in terms of distance between the head assigned by the parser and the actual head. Finally, we look at the error distribution in relation to sentence length. In the analysis, the results are aggregated over all the ten folds of the cross-validation.

8.1 Accuracy per Dependency Type

Table 7 gives the AS_U, labeled precision, labeled recall and labeled F-measure for individual dependency types. The table is sorted according to the AS_U results, and the average distance between head and dependent is given for each type.

We see that the parser cannot find labeled dependencies for the types that have fewer than 100 occurrences in the treebank, with the single exception of RELATIVIZER, which is the clitic "ki" (conjunction), written separately from the word it attaches to. Since this dependency type always occurs with the same particle, there is no sparse data problem.

If we exclude the low-frequency types, we can divide the results into three main groups. The first group consists of determiners, particles and nominal that have an AS_U score over 79% and link to nearby heads. The second group mainly contains subjects, objects and different kinds of adjuncts, with a score in the range 55–79% and

Table 7
Attachment score (AS_U), labeled precision (P), labeled recall (R) and labeled F measure for each dependency type in the treebank (n = count, dist = dependency length)

Label	n	$dist$	$\mathbf{AS}_U$	P	R	F
SENTENCE	7252	1.5	90.5	87.4	89.2	88.3
DETERMINER	1952	1.3	90.0	84.6	85.3	85.0
QUESTION.PARTICLE	288	1.3	86.1	80.0	76.4	78.2
INTENSIFIER	903	1.2	85.9	80.7	80.3	80.5
RELATIVIZER	85	1.2	84.7	56.6	50.6	53.4
CLASSIFIER	2048	1.2	83.7	74.6	71.7	73.1
POSSESSOR	1516	1.9	79.4	81.6	73.6	77.4
NEGATIVE.PARTICLE	160	1.4	79.4	76.4	68.8	72.4
OBJECT	7956	1.8	75.9	63.3	62.5	62.9
MODIFIER	11685	2.6	71.9	66.5	64.8	65.7
DATIVE.ADJUNCT	1360	2.4	70.8	46.4	50.2	48.2
FOCUS.PARTICLE	23	1.1	69.6	0.0	0.0	0.0
SUBJECT	4479	4.6	68.6	50.9	56.2	53.4
ABLATIVE.ADJUNCT	523	2.5	68.1	44.0	54.5	48.7
INSTRUMENTAL.ADJUNCT	271	3.0	62.7	29.8	21.8	25.2
ETOL	10	4.2	60.0	0.0	0.0	0.0
LOCATIVE.ADJUNCT	1142	4.2	56.9	43.3	48.4	45.7
COORDINATION	814	3.4	54.1	53.1	49.8	51.4
S.MODIFIER	594	9.6	50.8	42.2	45.8	43.9
EQU.ADJUNCT	16	3.7	50.0	0.0	0.0	0.0
APPOSITION	187	6.4	49.2	49.2	16.6	24.8
VOCATIVE	241	3.4	42.3	27.2	18.3	21.8
COLLOCATION	51	3.3	41.2	0.0	0.0	0.0
ROOT	16	-	0.0	0.0	0.0	0.0
Total	43572	2.5	76.0	67.0	67.0	67.0

a distance of 1.8–4.6 IGs to their head. This is the group where inflectional features are most important for finding the correct dependency. The third group contains distant dependencies with a much lower accuracy. These are generally relations like sentence modifier, vocative, and apposition, which are hard to find for the parser because they cannot be differentiated from other nominals used as subjects, objects or normal modifiers. Another construction that is hard to parse correctly is coordination, which may require a special treatment.

8.2 Error Distance

When we evaluate our parser based on the dependency direction, we obtain an AS_U of 72.2 for head-initial dependencies and 76.2 for head-final ones. Figure 10-a and Figure 10-b give the error distance distributions for head-initial and head-final dependencies based on the unlabeled performance of the parser. The x-axis in the figures gives the difference between indexes of the assigned head IG and the real head IG.

As stated previously, the head-initial dependencies constitute 5% of the entire dependencies in the treebank. Figure 10-a shows that for head-initial dependencies the parser has a tendency to connect the dependents to a head closer than the real head or in the wrong direction. When we investigate these dependencies, we see that 70.4% of them are connected to a head adjacent to the dependent and the parser finds 90.1%

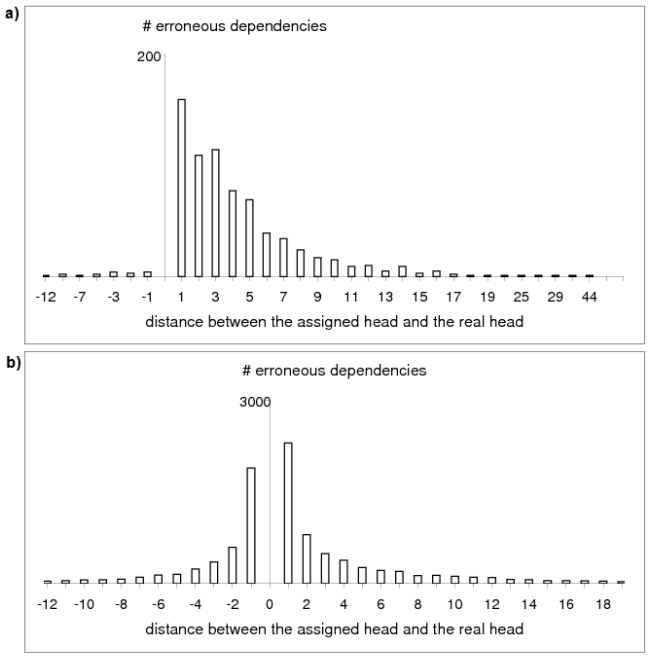


Figure 10
Error distance distributions a) for head-initial dependencies b) for head-final dependencies

of these dependencies correctly. Thus, we can say that the parser has no problem in finding adjacent head-initial dependencies. Moreover, 86.8% of the errors where the error distance is equal to 1 (Figure 10-a)[19] are due to the dependents being connected to the wrong IG of the correct head word. The error for finding the correct direction is 19.7% for these dependencies.

The parser is 100% successful in finding the direction of head-final dependencies. Furthermore, the errors that it makes while determining the correct head have a roughly normal distance distribution as can be seen from Figure 10-b.[20] We can see from the same figure that 57.3% of the errors fall within the interval of ± 2 IGs away from the actual head.

8.3 Sentence Length

Figure 11 shows the distribution of errors over sentences of different lengths. The x-axis plots sentence length (measured in number of dependencies), the y-axis shows the number of erroneous dependencies, and the z-axis indicates the frequency of a particular combination of error count and sentence length. As expected, the distribution is dominated by short sentences with few errors (especially sentences of up to seven dependencies with one error). The mean number of errors appears to be a linear function of sentence length, which would imply that the error probability per word does not increase with sentence length.

19 Meaning that the actual head and assigned head are adjacent.
20 Error distances with less than 40 occurrences are excluded from the figure.

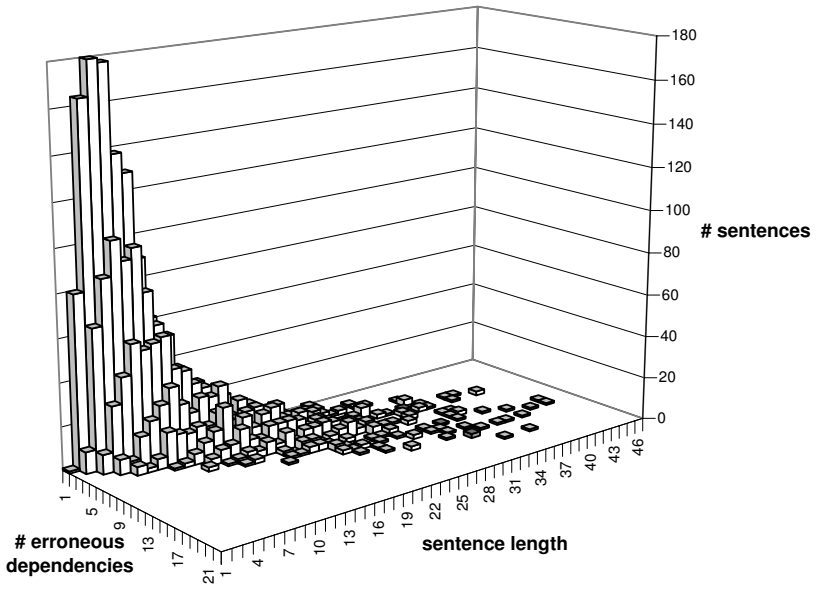


Figure 11
Error distribution in relation to sentence length

9. The Impact of Part-of-Speech Tagging

In all of the experiments reported above, we have used the gold-standard tags provided by the treebank. Another point that deserves investigation is therefore the impact of using tags automatically assigned by a part-of-speech tagger. With this purpose, we first used the two-level morphological analyzer of Oflazer (1994) to analyze all the words in the treebank[21] and then used the part-of-speech tagger of Yüret and Türe (2006), which has the best reported morphological disambiguator accuracy (96%) for Turkish. The complexity of morphological disambiguation in an agglutinative language like Turkish is due to the number of possible tags that can be assigned to a word (Yüret and Türe 2006). The number of potential morphological tags in Turkish is theoretically infinite due to productively derived forms.[22] The tagger should find the correct inflectional features and the IG structure as well as the correct part-of-speech categories, as shown in the following example:

kalemi
kale +Noun+A3sg+P1sg+Acc (*my castle* in accusative form)
kalem +Noun+A3sg+P3sg+Nom (*his pencil*)
kalem +Noun+A3sg+Pnon+Acc (*the pencil* in accusative form)

21 At the end of morphological analysis, it is seen that 39% of the words are ambiguous and 17% have more than two distinct morphological analyses.

22 The for treebank data, the number of distinct part-of-speech tags (defined as distinct combinantions of morphological features) is 718 for the word-based model of the classifier-based parser and 108 for the IG-based model.

When tested on our treebank data, the accuracy of the morphological disambiguator is 88.4%, including punctuation (which is unambiguous) and using a lookup table for the words that are not recognized by the morphological analyzer.[23] The lower accuracy of the morphological disambiguator on the treebank can be due to different selections in the annotation process of the morphological disambiguator training data (Yüret and Türe 2006) (which is totally different from the treebank data).

In order to investigate the influence of part-of-speech tagging errors, we used our best IG-based model and a lexicalized word-based model[24] with our classifier-based parser. One problem in the evaluation is that the assigned head IGs by the parser will not be relevant for some words where the part-of-speech tagger selects a morphological analysis with an IG structure totally different from the gold standard. There is no simple and straightforward solution to this problem, so we have used several evaluation metrics to get a more fine-grained picture of the impact of part-of-speech tagging. In all cases, WW_U scores only take into account whether the head word assigned to a dependent is correct or not, which means that any errors of the part-of-speech tagger can be ignored. Similarly, in calculating AS_U and AS_L scores for the word-based model, dependencies are assumed to be connected to the first IG of the head word without taking into consideration any errors in tags caused by the part-of-speech tagger. For the IG-based model, AS_U and AS_L are calculated as usual if both the dependent and the head word have exactly the same tag as in the gold standard. However, in the case of a tagging error in the dependent or head word (or both), dependencies have been scored according to four different models:

Default A dependency is scored as correct if it connects to the first IG of the correct head word (cf. the default assumption for word-based models).

HeadIG A dependency is scored as correct if it connects to the correct head word and the head IG has the same part-of-speech tag as in the gold standard.

BothIGs A dependency is scored as correct if it connects to the correct head word and both the dependent IG and the head IG have the same part-of-speech tag as in the gold standard.

BothWords A dependency is scored as correct if it connects to the correct head word and both the dependent word and the head word have the same IG segmentation and part-of-speech tags as in the gold standard.

Table 8 shows that the IG-based model and the word-based model are equally affected by the morphological disambiguation errors and have a drop in accuracy within similar ranges. (It can also be seen that, even with automatically tagged data, the IG-based model gives better accuracy than the word-based model.) Our most severe evaluation metric is *BothWords*, which penalizes all dependencies emanating from or landing on words that are incorrectly analyzed after morphological disambiguation (11.6% of all words). This metric is probably too severe, since we know that some of the errors in inflectional features do not affect the type of dependency very much. For example, if we put the adjective "küçük" (*small*) in front of the example given above (küçük kalemi), then the choice of morphological analysis of the noun has no impact on the fact that the adjective should be connected to the noun with dependency type

23 The words not recognized by the morphological analyzer are generally proper nouns, numbers and some combined words that are created in the development stage of the treebank and constitute 6.2% of the whole treebank. If these words are excluded, the accuracy of the disambiguator is 84.6%.

24 For this model, we added LEX features for σ_0,τ_0, τ_1 to the feature model of our word-based model in Table 4

Table 8
Impact of part-of-speech tagging

		AS_U	AS_L	WW_U
Word-based	*Gold standard*	*71.2*±0.3	*62.3*±0.3	*82.1*±0.9
	Tagged	69.5±0.3	59.3±0.3	80.2±0.9
IG-based	*Gold standard*	*76.0*±0.2	*67.0*±0.3	*82.7*±0.5
	Tagged Default	73.1±0.3	63.0±0.3	80.6±0.7
	Tagged HeadIG	73.3±0.3	63.2±0.3	80.6±0.7
	Tagged BothIGs	70.1±0.3	61.6±0.3	80.6±0.7
	Tagged BothWords	62.8±0.3	55.8±0.3	80.6±0.7

"MODIFIER". Moreover, most of the errors in part-of-speech categories will actually prevent the parser from finding the correct head word, which can be observed from the drop in WW_U accuracy (from 82.7 to 80.6 for the IG-based model). By contrast, the evaluation metric *HeadIG* ignores tagging errors on other IGs of the head word, as well as errors on the dependent word, which is reasonable given the assumption that dependencies always emanate from the last IG of the dependent word. Taking this as our IG-based metric, we can conclude that the use of an automatic morphological analyzer and disambiguator causes a drop in the range of 3 percentage points for unlabeled accuracy and 4 percentage points for labeled accuracy (for both word-based and IG-based models).

10. Related Work

The Turkish Treebank has recently been parsed by seventeen research groups in the CoNLL-X shared task on multilingual dependency parsing (Buchholz and Marsi 2006), where it was seen as the most difficult language by the organizers and most of the groups.[25] The following quotation is taken from Buchholz and Marsi (2006): "The most difficult data set is clearly the Turkish one. It is rather small, and in contrast to Arabic and Slovene, which are equally small or smaller, it covers 8 genres, which results in a high percentage of new FORM and LEMMA values in the test set."

The results for Turkish are given in Table 9. Our classifier-based parser obtained the best results for Turkish (with AS_U=75.8 and AS_L=65.7) and also for Japanese which is the only agglutinative and head-final language in the shared task other than Turkish (Nivre et al. 2006). The groups were asked to find the correct IG-to-IG dependency links. When we look at the results, we observe that most of the best performing parsers use one of the parsing algorithms of Eisner (1996), Nivre (2003) or Yamada and Matsumoto (2003) together with a learning method based on the maximum margin strategy. We can also see that a common property of the parsers which fall below the average (AS_L=55.4) is that they do not make use of inflectional features which is crucial for Turkish.[26]

25 The Turkish data used in the shared task is actually a modified version of the original treebank; some conversions are made on punctuation structures in order to keep consistency between all languages.
26 Actually, there are two parsers (Bick and Attardi) in this group which try to use parts of the inflectional features under special circumstances.

Table 9
CoNLL-X shared task results on Turkish

Teams	AS_U	AS_L
Nivre et al.	**75.8**	**65.7**
Johansson and Nugues	73.6	63.4
McDonald et al.	74.7	63.2
Corston-Oliver and Aue	73.1	61.7
Cheng et al.	74.5	61.2
Chang et al.	73.2	60.5
Yüret	71.5	60.3
Riedel et al.	74.1	58.6
Carreras et al.	70.1	58.1
Wu et al.	69.3	55.1
Shimizu	68.8	54.2
Bick	65.5	53.9
Canisius et al.	64.2	51.1
Schiehlen and Spranger	61.6	49.8
Dreyer et al.	60.5	46.1
Liu et al.	56.9	41.7
Attardi	65.3	37.8

11. Conclusion

In this article, we have investigated a number of issues in data-driven dependency parsing of Turkish. One of the main results is that IG-based models consistently outperform word-based models. This results holds regardless of whether we evaluate accuracy on the word level or on the IG level; it holds regardless of whether we use the probabilistic parser or the classifier-based parser; and it holds even if we take into account the problem caused by errors in automatic morphological analysis and disambiguation.

Another important conclusion is that the use of morphological information can increase parsing accuracy substantially. Again, this result has been obtained both for the probabilistic and the classifier-based parser, although the probabilistic parser requires careful manual selection of relevant features to counter the effect of data sparseness. A similar result has been obtained with respect to lexicalization, although in this case an improvement has only been demonstrated for the classifier-based parser, which is probably due to its greater resilience to data sparseness.

By combining the deterministic classifier-based parsing approach with an adequate use of IG-based representations, morphological information and lexicalization, we have been able to achieve the highest reported accuracy for parsing the Turkish Treebank.

Acknowledgments
We are grateful for the financial support from TUBITAK (The Scientific and Technical Research Council of Turkey) and Istanbul Technical University. We want to thank to the language technology group in Växjö University for providing us the classifier based parser platform (MaltParser) upon which we developed the methods in this article. We also want to thank Deniz Yüret for providing us his part-of-speech tagger and Eşref Adalı for his valuable comments.

References

Arun, Abhishek and Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL'05*, pages 302–313, Ann Arbor, MI.

Bikel, Daniel M. 2004. A distributional analysis of a lexicalized statistical parsing model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 182–189, Barcelona.

Bikel, Daniel M. and David Chiang. 2000. Two statistical parsing models applied to the Chinese treebank. In *Proceedings of the 2nd Chinese Language Processing Workshop*, pages 1–6, Hong Kong.

Black, Ezra, Frederick Jelinek, John D. Lafferty, David M. Magerman, Robert L. Mercer, and Salim Roukos. 1992. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 5th DARPA Speech and Natural Language Workshop*, pages 31–37, New York, NY.

Buchholz, Sabine and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 149–164, New York, NY. Association for Computational Linguistics.

Chang, Chih-Chung and Chih-Jen Lin, 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chung, Hoojung and Hae-Chang Rim. 2004. Unlexicalized dependency parser for variable word order languages based on local contextual pattern. In *Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 109–120, Seoul.

Collins, Michael. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of ACL'96*, pages 184–191, Santa Cruz, CA.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL'97*, pages 16–23, Madrid.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Collins, Michael, Jan Hajic, Lance Ramshaw, and Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of ACL'99*, pages 505–518, College Park, MD.

Corazza, Anna, Alberto Lavelli, Giorgio Satta, and Roberto Zanoli. 2004. Analyzing an Italian treebank with state-of-the-art statistical parsers. In *Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories*, pages 39–50, Tübingen.

Daelemans, Walter and Antal Vanden Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press, Cambridge.

Dubey, Amit and Frank Keller. 2003. Probabilistic parsing for German using sister-head dependencies. In *Proceedings of ACL'03*, pages 96–103, Sapporo.

Eisner, Jason. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 340–345, Copenhagen, 5-9 August.

Eryiğit, Gülşen, Joakim Nivre, and Kemal Oflazer. 2006. The incremental use of morphological information and lexicalization in data-driven dependency parsing. In *Proceedings of the 21st International Conference on the Computer Processing of Oriental Languages*, Singapore.

Eryiğit, Gülşen and Kemal Oflazer. 2006. Statistical dependency parsing of Turkish. In *Proceedings of EACL'06*, pages 89–96, Trento.

Gildea, Daniel. 2001. Corpus variation and parser performance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 167–202, Pittsburgh, PA.

Hajič, Jan, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Hladká. 2001. Prague dependency treebank 1.0 (final production label). CDROM CAT: LDC2001T10., ISBN 1-58563-212-0.

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

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL'03*, pages 423–430, Sapporo.

Kromann, Matthias T. 2003. The Danish dependency treebank and the underlying linguistic theory. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*, pages 217–220, Växjö.

Kudo, Taku and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 63–69, Taipei.

Levy, Roger and Christopher Manning. 2003. Is it harder to parse Chinese, or the Chinese treebank? In *Proceedings of ACL'03*, pages 439–446, Sapporo.

Magerman, David M. 1995. Statistical decision-tree models for parsing. In *Proceedings of ACL'95*, pages 276–283, Cambridge, MA.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Nilsson, Jens, Johan Hall, and Joakim Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proceedings of the 15th Nordic Conference of Computational Linguistics Special Session on Treebanks*, pages 119–132, Joensuu.

Nivre, Joakim. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies*, pages 149–160, Nancy.

Nivre, Joakim. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57, Barcelona.

Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Springer, Dordrecht.

Nivre, Joakim, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 49–56, Boston, MA.

Nivre, Joakim, Johan Hall, Jens Nilsson, Gülşen Eryiğit, and Stetoslav Marinov. 2006. Labeled pseudo-projective dependency parsing with support vector machines. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 221–225, New York, NY.

Nivre, Joakim and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of ACL'05*, pages 99–106, Ann Arbor, MI.

Oflazer, Kemal. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

Oflazer, Kemal. 2003. Dependency parsing with an extended finite-state approach. *Computational Linguistics*, 29(4):515–544.

Oflazer, Kemal, Bilge Say, Dilek Z. Hakkani-Tür, and Gökhan Tür. 2003. Building a Turkish treebank. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*. Kluwer, London, pages 261–277.

Ratnaparkhi, Adwait. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1–10, Providence, RI.

Sagae, Kenji and Alon Lavie. 2005. A classifier-based parser with linear run-time complexity. In *Proceedings of the 9th International Workshop on Parsing Technologies*, pages 125–132, Vancouver.

Sekine, Satoshi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2000. Backward beam search algorithm for dependency analysis of japanese. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 754–760, Saarbrücken.

Simov, Kiril, Gergana Popova, and Petya Osenova. 2002. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In Andrew Wilson, Paul Rayson, and Tony McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Lincom-Europa, Munich, pages 135–142.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. Springer, New York, NY.

Veenstra, Jorn and Walter Daelemans. 2000. A memory-based alternative for connectionist shift-reduce parsing. Technical Report ILK-0012, Tilburg University, Tilburg.

Yamada, Hiroyasu and Yuji Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proceedings of the 8th International Workshop on Parsing Technologies*, pages 195–206, Nancy.

Yüret, Deniz and Ferhan Türe. 2006. Learning morphological disambiguation rules for Turkish. In *Proceedings of HLT/NAACL'06*, pages 328–334, New York, NY.