The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish

Gülşen Eryiğit

Department of Computer Engineering Istanbul Technical University Istanbul, 34469, Turkey
gulsen.cebiroglu@itu.edu.tr

Abstract

The studies on dependency parsing of Turkish so far gave their results on the Turkish Dependency Treebank. This treebank consists of gold standard sentences where part-of-speech tags are manually assigned to each word and the words forming multi word expressions are also manually determined and combined into single units. For the first time, we investigate the results of parsing Turkish sentences from scratch and observe the accuracy drop at the end of processing raw data. We test one state-of-the art morphological analyzer together with two different morphological disambiguators. We both show separately the accuracy drop due to the automatic morphological processing and to the lack of multi word unit extraction. With this purpose, we use and present a new version of the Turkish Treebank where we detached the multi word expressions (MWEs) into multiple tokens and manually annotated the missing part-of-speech tags of these new tokens.

Keywords: Syntactic Parsing, Morphological Processing, Turkish

1. Introduction

Day by day, the amount of natural language data keeps increasing very rapidly by the high usage of social networks such as web forums, facebook and twitter. The growing need in semantic understanding of these written text in the real world applications (such as social CRMs, information retrieval systems and so on.) highlights the necessity and the importance of syntactic parsing. As the parsing community, we are used to evaluate our performances over treebanks' data where the previous nlp stages (e.g. tokenization, morphological analysis, morphological disambiguation, multi word expression extraction) before the parsing are assumed to be ideal, i.e., we are mostly focusing on the results of obtaining the exact sentence structure by using the manually tagged gold-standard words. However, we see that these automatic preprocessing stages are far from being perfect for many natural languages especially for the morphologically rich ones.

In recent studies, we see the growing interest of parsing raw data. Hogan et al. (2011), Eryiğit et al. (2011), Korkontzelos and Manandhar (2010) investigates the impact of detecting multi word expressions in parsing scores. Bengoetxea et al. (2011) investigates the impact of morphological disambiguation in dependency parsing of Basque. Lee et al. (2011) proposes a discriminative model for joint morphological disambiguation and dependency parsing for morphologically rich languages namely Czech, Latin, Ancient Greek and Hungarian.

This is the first study which makes an in depth investigation of the dependency parsing performance on raw Turkish data. [1] We are using a pipeline approach (Figure 1) where we first analyze the raw Turkish sentences by using the morphological analyzer of Oflazer (1994) and then we test two state-of-the-art morphological disambiguators (Yüret and Türe, 2006; Sak et al., 2008) in order to disambiguate multiple possible morphological analyses for each word. We then evaluate the parsing results by using a multilingual dependency parser (Nivre et al., 2007b).

Previous studies (Oflazer, 1994; Hakkani-Tür et al., 2002; Eryiğit et al., 2008; Buchholz and Marsi, 2006) deals with the morphologically rich and derivational structure of the Turkish language by representing the morphological information as inflectional groups (IGs) which are units smaller than words. A word in Turkish may consist one or more IGs and the syntactic dependencies are represented between these sub-units rather than words. This property of the language poses challenging problems during the extrinsic evaluation of the automatic morphological disambiguators on dependency parsing.

In this study, we mainly provide an intrinsic and extrinsic evaluation of the morphological preprocessing (analyzers&disambiguators) on the Turkish dependency parsing. This is a relatively complex work for Turkish since the automatically analyzed words may or may not have the same IG structure as in the gold-standard. We use a new version of the Turkish Dependency Treebank where the manually combined multi word expressions are firstly separated into multiple words. These words are then passed from the morphological analyzer and manually disambiguated (Eryiğit et al., 2011).[2] We make an in depth analysis of the problems encountered during the usage of the pipeline approach in order to parse raw data which we believe will be very useful for further studies in the field.

The paper is structured as follows: Section 2 will introduce some relevant properties of the Turkish language, Section 3 will present the used parsing framework, the data sets and the data preparation for different processing stages, Section

[1] A first attempt to parsing raw Turkish data has been made in Eryiğit et al. (2008) but this was rather a partial analysis; the authors has emphasized that the morphological analyzer has not produced any results for 6.2% of the words in the treebank and these were mainly multi word expressions expressed as single units.

[2] In this new version of the treebank, 2437 words' morphological analyses are manually disambiguated and the incoming and outgoing dependencies of the treebank sentences consisting these multi word expressions are rearranged.

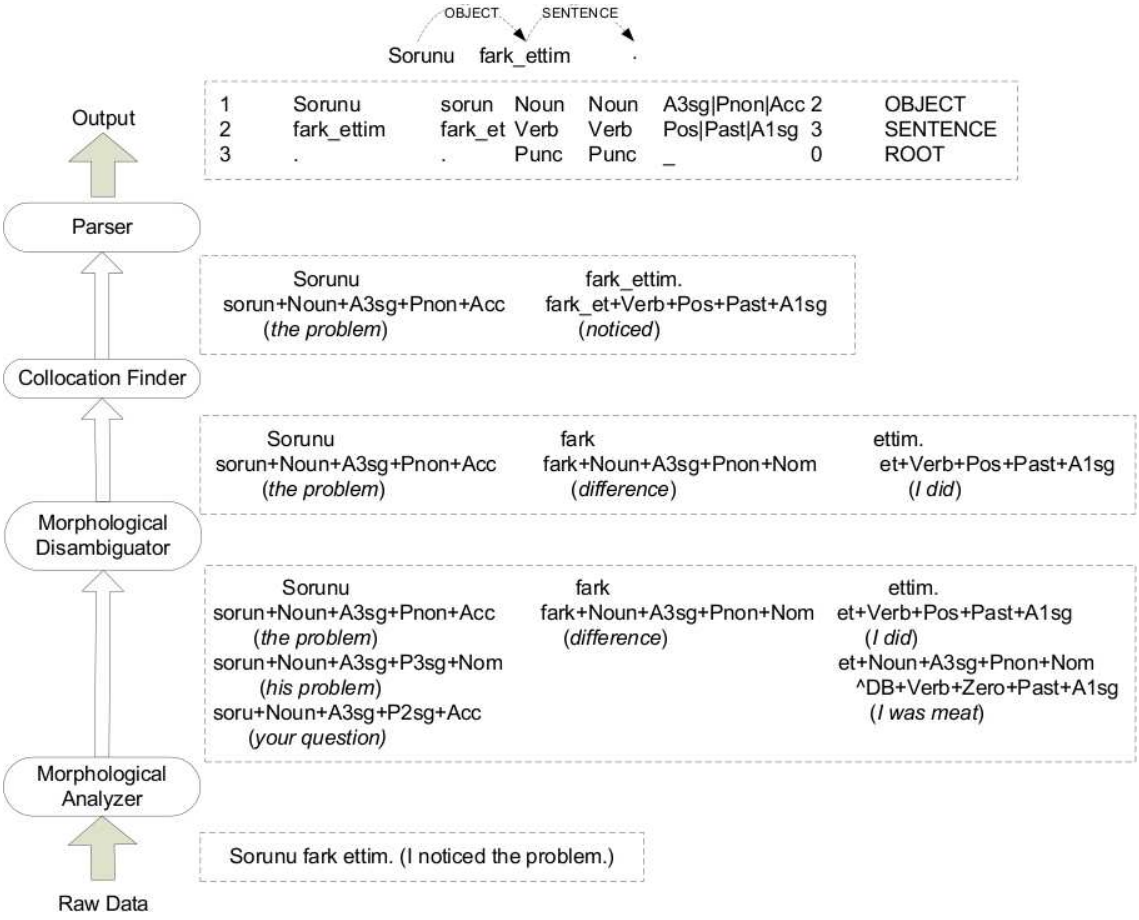

Figure 1: Flow of Turkish text processing

4 will give our evaluation strategy and the experimental results together with the discussions and finally Section 5 will make the conclusion.

2. Turkish

Turkish is an agglutinative language with a very rich morphological structure. This rich and derivational structure of the Turkish language is represented in the literature by the use of inflectional groups (IGs) which are units smaller than words. A word in Turkish may consist one or more IGs each consisting either a stem or a derivational suffix plus all the inflectional suffixes belonging to that stem/derivational suffix.

Figure 1 (Morphological Analyzer's output; the second layer, under each word) presents an example of possible morphological analyses for each of the three words in the given sentence. The word "ettim" has two possible morphological analyses: the first one "et+Verb+Pos+Past+A1sg" is the verb "to do" in past tense 1st singular form and the second one "et+Noun+A3sg+Pnon+Nom +ˆDB +Verb+Zero+Past+A1sg" is an analysis which consists of two IGs seperated by a derivational boundary (ˆDB). The first IG is the noun "meat" in singular nominative form and the second IG which is derived from the first one is the verb "to be meat" in past tense 1st singular form.

The morphological disambiguation level gives the correct analysis of each word in the given context. The combination of MWEs into single units is shown at the third level of the figure. The last layer shows the dependency structure of the sentence. In dependency relations, the head of a whole word is not just another word but a specific IG of another word. Figure 2 shows the phenomenon in the simple sentence: "küçük odadayım" (*I'm in the small room*). The word "odadayım" (*I'm in the room*) is formed from two IGs; the verb "being in the room" is derived from the inflected noun "odada" (*in the room*). In this example, the adjective "küçük" (*small*) should be connected to the first IG of the second word. It is the word "oda" (*room*) which is modified by the adjective "small", not the derived verb form "odadayım" (*I'm in the room*). Thus, both the correct head word and the correct IG in the head word should be determined by the dependency parser.

3. Configuration

In our experiments we are using four analyzers for three different processing layers:

1. A two-level morphological analyzer (Oflazer, 1994)
2. Two morphological disambiguators (Yüret and Türe, 2006; Sak et al., 2008)
3. A data-driven dependency parser (Nivre et al., 2007b)

We do not use any automatic multi word expression extraction software to test with since to our knowledge there is

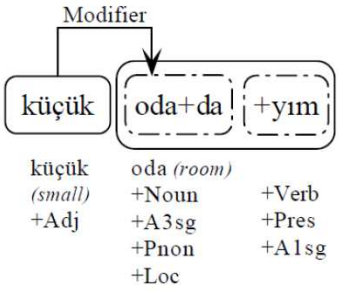

Figure 2: Dependency relations
A1sg = 1sg number/person agreement, A3sg = 3sg number/person agreement, Loc = Locative case, Pnon = No possessive agreement, Pres = Present Tense

not any high performance extractors for Turkish yet. For this reason, in this study we prefer to show the performance drop due to the lack of MWE extraction. One should refer to Eryiğit et al. (2011) for further analysis and the increase that could be obtained by an ideal MWE analyzer.
We are using MaltParser v1.5.1 (Nivre et al., 2007b) which is a data-driven dependency parser whose success is reported to be very high across a variety of different languages (Nivre et al., 2006). The parser's current version uses a support vector machine (SVM) classifier for predicting the parser's actions. For the repeatability of the results we used exactly the same feature representation and parser options from Eryigit et al. (2008) . The cited reference gives these options in details so we do not repeat them here again. As before, we train the parser with the original treebank. But in the following sections we will provide two results with a single difference in SVM options. One by using the entire training data at a time during the training of the SVM and one where the training data has been divided into smaller sets (based on the major part of speech category of the next token in the queue) as a way of reducing SVM training times without a significant decrease in accuracy (Nivre et al., 2006). We provide the former results to show the maximum performance that could be obtained by using the gold-standard data.

3.1. Data Sets

We use the METU-Sabanci Turkish Treebank (Oflazer et al., 2003) of 5635 sentences and ITU validation set (Eryiğit, 2007) of 300 sentences for our tests. Both of these treebanks uses the dependency formalism in order to represent the syntax relations within a sentence. The words are tagged with gold-standard part-of-speech tags. Multi word expressions are expressed as a single unit in syntax relations. These MWEs' are tagged only with the part-of-speech tag and inflectional features of the last word conforming that MWE.
In order to be able to test the parsing performance on raw data, we are using a new version of the Turkish Treebank[3] (Eryiğit et al., 2011). In this version, the MWEs are first split into separate words and the new words coming out after this splitting (the ones except the last word confirming that MWE) which were left with unassigned part-of-speech tags, are first passed from the morphological analyzer. The possible morphological outputs are then manually disambiguated. In our experiments we will refer to this version as Vd (standing for Version detached). These words are then linked to each other in the dependency representation with a new dependency label called "MWE". For further sections, one should keep in mind that this new label doesn't exist in the original treebank.
The treebank versions which will be referred during the experiments are as follows:

- Vo: Version Original
- Vd: Version Detached
- Vraw: Version Raw

The latter one (Vraw) is an automatically created version by first taking the words from Vd, passing them from the automatic morphological analysis and then disambiguation. The data is then automatically converted to the treebank format in order to be processed with the syntactic parser. Some conversions made during this process is explained in the following subsection (Subsection 3.2.).

3.2. Data Preparation

In our experiments, although we used the same morphological analyzer which is used in the construction phase of the treebank, we noticed that the output tag set of the morphological analyzer differs slightly from the tag set which is used in the treebanks. The matching between the tag sets is straightforward and may be realized easily by writing conversion scripts defined in the following paragraphs. Since the disambiguators were trained with the original tags produced by the morphological analyzer, these conversions are made after the disambiguation stage, during the conversion of the selected morphological analyses to the treebank format. The intrinsic evaluation of the disambiguators are made after these conversions by comparing with the gold standard treebank analyses. Thus, some of the conversions described in this section will have a positive influence on their performance evaluation.
An example of the tag differences is the following: The adverbs are tagged as "Adverb" in the morphological analyzer output but this tag is abbreviated as "Adv" in the treebank data. Another example is: the tag for adjectives with future participle form is represented with a single tag "AFutPart" in the treebank instead of "Adj+FutPart" in the morphological analyzer output. The full list of such dissimilarities is given in Table 1.
In addition to these, during data preparation, some obvious and very frequent bugs of the morphological analyzer is corrected automatically, e.g. the most frequent meaning of the word "ise" in Turkish is "*if*" but the analysis (ise+Conj) referring to this meaning is never produced with the morphological analyzer in hand. Such problems were also occurring for some punctuations, numerical expressions where the analyzer was not able to produce any result at all. 397/427 of unassigned tags were punctuations

[3]The new treebank is available from `http://web.itu.edu.tr/gulsenc/resources.htm`

```
if(dependency is connected to the correct word){
        unlabeled_correct_word_count++;
        if(dependency label is correct)
                labeled_correct_word_count++;
        if(the connected word has totally the same morphological
                analysis with the gold standard){
                if(dependency is connected to the correct IG){
                        unlabeled_correct_IG_count++;
                        if(dependency label is correct)
                                labeled_correct_IG_count++;
                }
        }
        else{
                if(the connected IG has the same main POS tag
                         with the gold standard)
                        unlabeled_correct_IG_count++;
                        if(dependency label is correct)
                                labeled_correct_IG_count++;
        }
}
```

Figure 3: Evaluating the dependency links for automatically produced tags

Morph. Output	Treebank
"Adverb"	"Adv"
"Inf1" or "Inf2" or "Inf3"	"Inf"
"Pron"+X (X is either "Demons" or "Ques" or "Pers" or "Reflex")	X'P'
"Adj"+X (X is either "FutPart" or "PastPart" or "PresPart" or "Inf")	'A'X
"WithoutBeingAbleToHaveDoneSo"	"WithoutHaving DoneSo"

Table 1: Distinct Tags appearing in the Morphological Analyzer's Output and the Treebanks

and numbers. The remaining 30 words to which the morphological analyzer couldn't assign any tag are tagged as proper nouns.

Finally, the morphologically analyzed sentences are converted to the CoNLL format where each IG is represented as a separate unit.

4. Experiments

4.1. Evaluation Metrics

As opposed to other languages, in recent studies (Buchholz and Marsi, 2006; Nivre et al., 2007a) the evaluation of dependency structures for the Turkish language is not based on the calculation of correct dependencies betweens "words". As explained in Section 2., since the dependencies are shown between the inflectional groups of the dependent and head words, the evaluation is also based on the correct dependencies between the correct IGs. The main evaluation metrics that we use are the unlabeled attachment score (AS_U) and labeled attachment score (AS_L), namely, the proportion of IGs 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. In addition to these, 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. For the experiments on METU-Sabanci Turkish Treebank, we report the results as mean scores of the ten-fold cross-validation, together with standard error. We also provide the results on the ITU validation set by using a model trained with the entire sentences of the METU-Sabanci Turkish Treebank. Following previous studies, the dependencies emanating from punctuations are excluded from the evaluation.

4.2. Evaluation strategy for automatically analyzed tokens

The automatic morphological analysis process described in Section 3. (the first two layers) may or may not produce the same IG structure with the gold standard annotation for a specific word in the treebank. As an example, the analyzers may erroneously select the second analysis[4] for the word "ettim" with two IGs instead of the first one[5] with a single IG (Figure 1). In that case, evaluating a dependency link as correct or faulty becomes a problematic duty. There are many possible approaches that may be adopted ranging from too mild to too severe. Most of these are discussed in Eryiğit et al. (2008). In this study, we adopt the strategy whose pseudocode is given in Figure 3. This approach ignores the IG structure of the dependent word whether it is assigned correctly or not. This is because the dependent is always connected with its last IG. For the head word, if the assigned morphological analysis is exactly the same with the gold standard, then the evaluation is as explained in Section 4.1.. In the case that the morphological analysis

[4] "et+Noun+A3sg+Pnon+Nom+DB+Verb+Zero+Past+A1sg"

[5] "et+Verb+Pos+Past+A1sg"

		Yüret and Türe (2006)	Sak et al. (2008)
Reported Accuracy	including punctuations	95.82	96.45
Calculated Accuracy on METU-Sabancı Treebank	including punctuations	78.76	87.67
	excluding punctuations	73.96	84.89
Calculated Accuracy on ITU Validation Set	including punctuations	78.23	87.84
	excluding punctuations	74.87	86.09

Table 2: Intrinsic Evaluation of the Morphological Disambiguators

	METU-Sabancı			ITU		
tested on	AS_U	AS_L	WW_U	AS_U	AS_L	WW_U
Eryiğit et al. (2008) on Vo	76.0±0.2	67.0±0.3	82.7±0.5	x	x	x
Eryiğit et al. (2008) on partially raw data	73.3±0.3	63.2±0.3	80.6±0.7	x	x	x
Vo (SVM trained with the whole data)	76.1±0.2	67.4±0.3	83.0±0.2	79.90	71.98	84.25
Vo	75.9±0.2	67.0±0.2	82.3±0.2	80.45	71.82	84.22
Vd	74.4±0.2	62.8/66.0±0.3	80.9±0.2	80.08	69.84/71.89	83.88
Vraw disamb. with Sak et al. (2008)	70.7±0.2	58.3/61.1±0.2	78.5±0.2	75.53	64.24/66.17	80.49
Vraw disamb. with Yüret and Türe (2006)	66.1±0.3	51.3/53.8±0.2	74.9±0.3	70.34	55.23/56.97	76.95

Table 3: Parsing performance (trained on the original treebank (Vo))

is different than the gold standard then a dependency link is accepted to be correct if the dependent is connected to the correct head word and the head IG has the same main part-of-speech category as the gold-standard.

4.3. Results

We first measure the coverage of the automatic morphological analyzer. Since the analyses produced by the morphological analyzer do not always cover the perfect analysis, we observe that during the creation of the treebank (Oflazer et al., 2003) there exist some manipulation over these results; the human annotators did not always select an analysis from the automatically produced ones but instead, they sometimes added a new analysis. As an example, the word "o" ("*that*") is either a pronoun or a determiner in Turkish. We observe that the analyzer sometimes assigns a wrong tag such as "Noun" to this word when it is written with a capital letter. We first measured the coverage of the morphological analyzer (the percentage of the gold standard tag to be one of the analyses produced by the analyzer) as 97.8%. That is, in 2.2% of the cases, the gold-standard tag doesn't exist within the produced possible analyses.

As the second step, we make an intrinsic evaluation of our two morphological disambiguators. We used the disambiguators with their pretrained models. Table 2 shows the performances of the disambiguators on the METU-Sabancı Treebank and ITU Validation Set as well as their reported accuracies in related articles (Yüret and Türe, 2006; Sak et al., 2008). Since the punctuations are excluded during the evaluation of the parsing performance, here, we provide the results both by including and excluding punctuations. In this evaluation, a selection of a disambiguator is accepted to be true if the selected morphological analysis has exactly the same structure with the gold standard analysis; the same POS tags and inflectional features at IG-level. From the results, we observe that the accuracies are far from the reported values on the manually annotated gold-standard treebank data. The disambiguator of Sak et al. (2008) looks to perform better when compared to the disambiguator of Yüret and Türe (2006). The reason of this difference looks like the tendency of the latter one to prefer/assign tags with higher number of inflectional features; the average number of IGs per words in the original treebank is 1.25, where as this number is 1.23 in the output of Sak et al. (2008) and 1.33 in Yüret and Türe (2006).

In order to see the effect of the automatic morphological analysis on the dependency parsing performance, our third set of experiments are given in Table 3. The first two lines of the table give the results reported in the study of Eryiğit et al. (2008). The first line gives the parsing accuracy on the original treebank whereas the second line reports the accuracy on partially raw data [1] .

We first replicate this study with the new Maltparser version and give our results on the third and fourth lines. As explained in Section 3., in order to see the highest performance, we first use the default SVM options where the entire data is used at a time during the training. These results are given in the 3^{rd} line "Vo (SVM trained with the whole data)". The remaining of the table are the tests conducted by using extra SVM splitting options (Section 3.). At the 4^{th} line, we notice the slight decrease in accuracy using this strategy.

Our results on Vd (where MWEs are split into multiple units in test data), the 5^{th} line, show the decrease in accuracy with the lack of multiword extraction. The 6^{th} and 7^{th} lines gives the results obtained when the test data is morphologically analyzed by using the previously introduced mechanism. The dependencies at the output of these tests are compared with the gold standard dependencies from Vd. We observe that we again obtain better results by using the disambiguator of Sak et al. (2008). We notice an accuracy decrease of 5.2 in AS_U score and 3.8 in WW_U score. Nearly 1.5 of this drop is due to the lack of MWE extraction and the remaining part is due to the errors caused by automatic morphological analysis.

Following Eryiğit et al. (2011), for the AS_L scores we

provide two values: the first one is the normal evaluation which is a very strict one since the parser is unaware of the "MWE" dependency type (Section 3.1.) during the training, the second one is a much lighter evaluation where the dependencies of type "MWE" is accepted to be correct if the head and dependent is determined correctly. We again notice a nearly 6 point decrease in this scores (67.0 -> 61.1, 67.0->58.3).

The second part of the table also gives the evaluations over the ITU validation set. Again, there is a performance drop of nearly 5 percentage points in all scores. The drop due to MWE extraction in these tests are rather smaller due to the smaller size of test data and smaller number of MWEs accordingly.

5. Conclusion

In this study, we made an in depth analysis on the impact of automatic morphological analysis on parsing raw Turkish data. We provided the mechanism to use with the existing tools and the conversions to be made for their integration. Since we still used properly written sentences (chosen and corrected for treebank creation), we may still not claim that our results reflect the performance on purely raw data such as spoken language or web data. Further analysis should be made on such data and new algorithms should be designed for increasing the parsing performance for Turkish which is still very low when compared to other languages.

Acknowledgments

The author wants to thank Tugay İlbay and Ozan Arkan Can for their help during the manual annotation of the new treebank version.

6. References

Kepa Bengoetxea, Arantza Casillas, and Koldo Gojenola. 2011. Testing the effect of morphological disambiguation in dependency parsing of Basque. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 28–33, Dublin, Ireland, October. Association for Computational Linguistics.

Sabine Buchholz 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.

Gülşen Eryiğit, Tugay Ilbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 45–55, Dublin, Ireland, October. Association for Computational Linguistics.

Gülşen Eryiğit. 2007. ITU validation set for Metu-Sabancı Turkish treebank.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

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

Deirdre Hogan, Jennifer Foster, and Josef van Genabith. 2011. Decreasing lexical data sparsity in statistical syntactic parsing - experiments with named entities. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 14–19, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 636–644, Los Angeles, California, June. Association for Computational Linguistics.

John Lee, Jason Naradowsky, and David A. Smith. 2011. A discriminative model for joint morphological disambiguation and dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 885–894, Portland, Oregon, USA, June. Association for Computational Linguistics.

Joakim Nivre, 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.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007a. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Stetoslav Marinov, and Erwin Marsi. 2007b. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal*, 13(2):99–135.

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

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

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *GoTAL 2008*, volume 5221 of *LNCS*, pages 417–427. Springer.

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