ITU Validation Set for Metu-Sabancı Turkish Treebank

Gülşen Eryiğit[*]
Istanbul Technical University

This paper introduces a validation set for the available dependency treebank of Turkish. The main aim of the preparation of this dataset is to serve as the test set of the CoNLL-XI shared task (shared task of the Conference on Computational Natural Language Learning 2007) and the data will be available from the webpage `http://www3.itu.edu.tr/~gulsenc/treebank` *after the shared task.*

1. Introduction

The Turkish Treebank (Oflazer et al. 2003; Atalay, Oflazer, and Say 2003) created by the Middle East Technical University and Sabancı University is available to the researchers since 2003 and it is used by many researchers since then (Eryiğit and Oflazer 2006; Eryiğit, Nivre, and Oflazer 2006; Eryiğit, Adalı, and Oflazer 2006; Nivre et al. 2007; Çakıcı and Baldridge 2006; Buchholz and Marsi 2006; Yüret 2006; Wu, Lee, and Yang 2006; Dreyer, Smith, and Smith 2006; Shimizu 2006; Schiehlen and Spranger 2006; Riedel, Çakıcı, and Meza-Ruiz 2006; Johansson and Nugues 2006; McDonald, Lerman, and Pereira 2006; Liu et al. 2006; Chang, Do, and Roth 2006; Corston-Oliver and Aue 2006; Cheng, Asahara, and Matsumoto 2006; Carreras, Surdeanu, and Marquez 2006; Canisius et al. 2006; Bick 2006; Attardi 2006; Eryiğit 2006). Although it has some inconsistencies and still continues to be updated with newer versions[1], it served very much in the recent years for the development of the research on dependency parsing of Turkish.

The Turkish treebank is composed of 5635 sentences and annotated with dependency structures. The modest data size of the treebank has been mentioned in many studies (Nivre et al. 2007; Buchholz and Marsi 2006). There is no need to say that the size should be increased for better research on the field, but we should also state that the small size of the number of words (48K) of this treebank can be actually related to one of the features of the language itself. In the treebank, the average number of words in a sentence is 8.6 which is very lower when compared to other languages. This is since in Turkish, the words are sometimes equivalent to a whole sentence in another language which is a result of its agglutinative structure. This property of the language makes look the treebank smaller than it is when compared to the other treebanks having similar number of sentences (refer to Nivre et al. (2007) for further analysis).

This paper presents the validation set prepared at Istanbul Technical University (ITU) for the Turkish Treebank. We adopted the same annotation scheme with the original treebank and annotated the sentences with dependency structures. The remaining of the paper first presents the structure of the prepared dataset (Section 2), then its

[*] Department of Computer Engineering, Istanbul Technical University, 34469 Istanbul, Turkey. E-mail: gulsen.cebiroglu@itu.edu.tr

[1] The changes between the versions of the treebank have been explained in Eryiğit (2006).

available data formats (Section 3) and finally its differences from the previous versions of the treebank (Section 4).

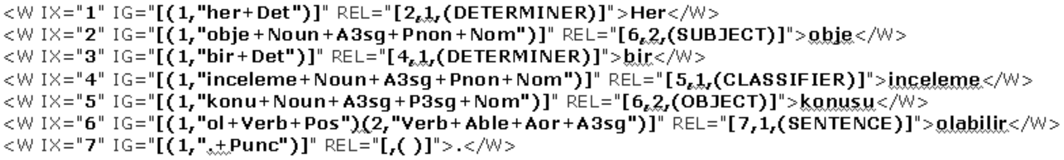
2. Validation Set

ITU Validation Set contains 300 sentences from 3 different genres (20% article, 20% novels and 60% short stories). The sentences are first analyzed with the morphological analyzer of Oflazer (1994) and then multiple morphological analyses are manually disambiguated. The sentences are then manually annotated according to dependency structure. Two annotators worked during the preparation of the dataset. Since, most of the observed inconsistencies on the current treebank is due to the incoherence between different annotators, during the preparation of the validation set the annotators were charged with different stages of the annotation process; the sentences are first morphologically disambiguated by one annotator then the second annotator double-checked the results of this disambiguation phase and annotated the dependencies simultaneously. We believe that this working style resulted in a viable validation set.

The dependency annotator used a special dependency type to emphasize the collocation structures. We then automatically combined these collocations[2] into single units and reindex the sentences by using scripts.

3. Data formats

The validation set is available in two different data formats[3]: *XML Data Format* which is the Turkish treebank original data format and *Conll Data format* which is the data format used in the Conll-X (Shared task on on Multi-lingual Dependency Parsing) and Conll-XI (Multilingual Track of the shared task). Please refer to Say (2004) and Buchholz and Marsi (2006) for the details of these formats. Figure 1[4] and Figure 2 give the representation of the sentence "Her obje bir inceleme konusu olabilir." *(Each object can be an investigation topic)* with these data formats.



```
<W IX="1" IG="[(1,"her+Det")]" REL="[2,1,(DETERMINER)]">Her</W>
<W IX="2" IG="[(1,"obje+Noun+A3sg+Pnon+Nom")]" REL="[6,2,(SUBJECT)]">obje</W>
<W IX="3" IG="[(1,"bir+Det")]" REL="[4,1,(DETERMINER)]">bir</W>
<W IX="4" IG="[(1,"inceleme+Noun+A3sg+Pnon+Nom")]" REL="[5,1,(CLASSIFIER)]">inceleme</W>
<W IX="5" IG="[(1,"konu+Noun+A3sg+P3sg+Nom")]" REL="[6,2,(OBJECT)]">konusu</W>
<W IX="6" IG="[(1,"ol+Verb+Pos")(2,"Verb+Able+Aor+A3sg")]" REL="[7,1,(SENTENCE)]">olabilir</W>
<W IX="7" IG="[(1,".+Punc")]" REL="[,( )]">.</W>
```

Figure 1
XML Data Format

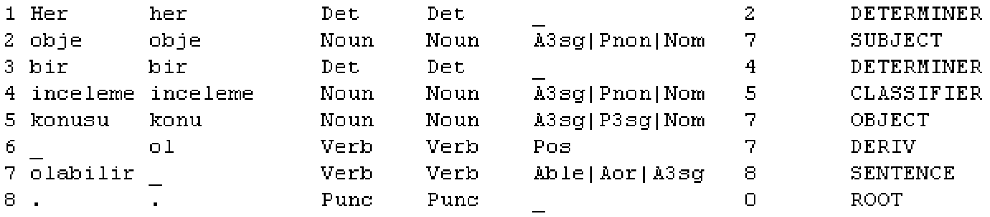
4. Differences from the previous versions

The recent official version of the Turkish treebank is the version used in the Conll-X shared task (Buchholz and Marsi 2006). This version is available as two subversions (one in XML and one in Conll format) from the treebank website `http://www.ii.`

[2] In the treebank, the words in a collocation have been combined into single units by putting an underscore "_" character in between.

[3] Actually, it is prepared in the original treebank XML format and then converted to Conll format.

[4] The fields "Lem" and "Morph", which are originally available in the treebank format but are empty in its current state, are removed from the figure because of the space limit.



```
1 Her       her        Det    Det    _               2    DETERMINER
2 obje      obje       Noun   Noun   A3sg|Pnon|Nom   7    SUBJECT
3 bir       bir        Det    Det    _               4    DETERMINER
4 inceleme  inceleme   Noun   Noun   A3sg|Pnon|Nom   5    CLASSIFIER
5 konusu    konu       Noun   Noun   A3sg|P3sg|Nom   7    OBJECT
6 _         ol         Verb   Verb   Pos             7    DERIV
7 olabilir  _          Verb   Verb   Able|Aor|A3sg   8    SENTENCE
8 .         .          Punc   Punc   _               0    ROOT
```

Figure 2
Conll Data Format

metu.edu.tr/~corpus/corpus.html. There is one major difference between these two subversions. The data used in the Conll-X shared task (in Conll format) is actually a variant of the treebank in XML format; some conversions are made on punctuation structures in order to keep consistency between all languages[5]. In Conll-XI, the entire treebank will be used as the training data and the validation set introduced in this paper will be used as the test data.

The treebank which will be used this year differs from the previous year mainly in two points:

- Unlike to Conll-X, for Conll-XI shared task, no conversion is applied to the punctuation structures,
- All the dependencies emanating from and coming to the words with a special stem "değil"[6] have been re-annoted in order to keep consistency on the overall treebank.

Following the changes in the treebank, the validation set is also prepared according to the final structure of the treebank and differs from Conll-X Turkish data and the original treebank on the items listed above.

5. Conclusion

In this paper, a validation set of 300 sentences for the Turkish Treebank has been introduced. The data set has been prepared according to the same annotation style of the original treebank and will be publicly available after the Conll-XI shared task from http://www3.itu.edu.tr/~gulsenc/treebank. We aim to improve the size of the data as future work.

Acknowledgments

The author wants to thank to Prof. Kemal Oflazer for his valuable comments on the development of the validation set and Prof. Joakim Nivre for discussions on the Turkish Treebank.

References

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

[5] refer to http://nextens.uvt.nl/~conll/software.html#conversion for further discussion

[6] This is a special word which occurs under different part-of-speech categories (Verb and Conj). The annotation manner for this verb is modified in the new version of the treebank.

Attardi, Giuseppe. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of CONLL-X*, pages 166–170, New York.

Bick, Eckhard. 2006. Lingpars, a linguistically inspired, language-independent machine learner for dependency treebanks. In *Proceedings of CONLL-X*, pages 171–175, New York.

Buchholz, Sabine and Erwin Marsi. 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of CONLL-X*, pages 149–164, New York.

Çakıcı, Ruket and Jason Baldridge. 2006. Projective and non-projective Turkish parsing. In *Proceedings of the 5th International Treebanks and Linguistic Theories Conference*, pages 43–54, Prague.

Canisius, Sander, Toine Bogers, Antal van den Bosch, Jeroen Geertzen, and Erik Tjong Kim Sang. 2006. Dependency parsing by inference over high-recall dependency predictions. In *Proceedings of CONLL-X*, pages 176–180, New York.

Carreras, Xavier, Mihai Surdeanu, and Lluis Marquez. 2006. Projective dependency parsing with perceptron. In *Proceedings of CONLL-X*, pages 181–185, New York.

Chang, Ming-Wei, Quang Do, and Dan Roth. 2006. A pipeline model for bottom-up dependency parsing. In *Proceedings of CONLL-X*, pages 186–190, New York.

Cheng, Yuchang, Masayuki Asahara, and Yuji Matsumoto. 2006. Multi-lingual dependency parsing at naist. In *Proceedings of CONLL-X*, pages 191–195, New York.

Corston-Oliver, Simon and Anthony Aue. 2006. Dependency parsing with reference to Slovene, Spanish and Swedish. In *Proceedings of CONLL-X*, pages 196–200, New York.

Dreyer, Markus, David A. Smith, and Noah A. Smith. 2006. Vine parsing and minimum risk reranking for speed and precision. In *Proceedings of CONLL-X*, pages 201–205, New York.

Eryiğit, Gülşen. 2006. *Türkçenin Bağlılık Ayrıştırması (Dependency Parsing of Turkish)*. Ph.D. thesis, Istanbul Technical University, Istanbul.

Eryiğit, Gülşen, Eşref Adalı, and Kemal Oflazer. 2006. Türkçe cümlelerin kural tabanlı bağlılık analizi (Rule-based dependency parsing of Turkish sentences). In *Proceedings of the 15th Turkish Symposium on Artificial Intelligence and Neural Networks*, pages 17–24, Muğla.

Eryiğit, Gülşen, Joakim Nivre, and Kemal Oflazer. 2006. The incremental use of morphological information and lexicalization in data-driven dependency parsing. *Computer Processing of Oriental Languages, Beyond the Orient: The Research Challenges Ahead, Springer*, LNAI 4285:498–507.

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

Johansson, Richard and Pierre Nugues. 2006. Investigating multilingual dependency parsing. In *Proceedings of CONLL-X*, pages 206–210, New York.

Liu, Ting, Jinshan Ma, Huijia Zhu, and Sheng Li. 2006. Dependency parsing based on dynamic local optimization. In *Proceedings of CONLL-X*, pages 211–215, New York.

McDonald, Ryan, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CONLL-X*, pages 216–220, New York.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Stetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal*, 13(1):1–41.

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

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.

Riedel, Sebastian, Ruket Çakıcı, and Ivan Meza-Ruiz. 2006. Multi-lingual dependency parsing with incremental integer linear programming. In *Proceedings of CONLL-X*, pages 226–230, New York.

Say, Bilge. 2004. Metu-sabancı turkish treebank user guide.

Schiehlen, Michael and Kristina Spranger. 2006. Language independent probabilistic context-free parsing bolstered by machine learning. In *Proceedings of CONLL-X*, pages 231–235, New York.

Shimizu, Nobuyuki. 2006. Maximum spanning tree algorithm for non-projective labeled dependency parsing. In *Proceedings of CONLL-X*, pages 236–240, New York.

Wu, Yu-Chieh, Yue-Shi Lee, and Jie-Chi Yang. 2006. The exploration of deterministic and efficient dependency parsing. In *Proceedings of CONLL-X*, pages 241–245, New York.

Yüret, Deniz. 2006. Dependency parsing as a classification problem. In *Proceedings of CONLL-X*, pages 246–250, New York.