ITU Validation Set
for Metu-Sabancı Turkish Treebank

Gülşen ERYİĞİT
gulsen.cebiroglu@itu.edu.tr;

Tuğba PAMAY
pamay@itu.edu.tr;

ABSTRACT

This paper presents the ITU Turkish Dependency Validation Set firstly introduced in 2007 [36] in order to serve as the test set of the CoNLL-XI shared task (shared task of the Conference on Computational Natural Language Learning 2007 [28]). The dataset is available from http://web.itu.edu.tr/gulsenc/treebanks.html and is used by several academic studies so far.

1 Introduction

The Turkish Treebank [1], [20] created by the Middle East Technical University and Sabanci University is available to the researchers since 2003 and it is used by many researchers since then [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [22], [24], [25], [26], [27]. 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 [19], [4]. 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 low 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 [19] for further analysis).

This paper presents the validation set prepared at Istanbul Technical University (ITU) for the Turkish Treebank. The same annotation scheme with the original treebank has been adapted and the sentences are annotated with dependency structures. The presented language resource "ITU Validation Set" which is firstly introduced and used in Conll-XI [27] has been used in many other studies so far. Some of which are [28], [29], [30], [31], [32], [33], [34], [35] . The remaining of the paper first presents the structure of the prepared dataset (Section 2), then its 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 [21] 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

[1] The changes between the versions of the treebank have been explained in [15].

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 [23] and [4] 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.

4 Differences from the previous versions

The recent official version of the Turkish treebank is the version used in the Conll-X shared task [4]. This version is available as two subversions (one in XML and one in Conll format)
from the treebank website http://www.ii.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 below.

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 is publicly available from *http://web.itu.edu.tr/gulsenc/treebanks.html*.

6 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.

[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.
[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.

```
<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

```
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

7 References

[1] **Atalay, N. B., Oflazer, K.,** and **Say, B..** 2003. The annotation process in the turkish treebank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*.

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

[3] **Bick, E..** 2006. LingPars, a Linguistically Inspired, Language-Independent Machine Learner for Dependency Treebanks. In *Proceedings of CONLL-X, pages 171-175, New York.*

[4] **Buchholz, S.,** and **Marsi, E..** 2006. Conll-X shared task on multilingual dependency parsing. In *Proceedings of CONLL-X*, pages 149-164, New York.

[5] **Çakıcı, R.,** and **Baldridge, J..** 2006. Projective and Non-Projective Turkish Parsing. In *Proceedings of the 5th International Treebanks and Linguistic Theories Conference*, pages 43-54, Prague.

[6] **Canisius, S., Bogers, T., Bosch van de, A., Geertzen, J.,** and **Tjong Kim Sang, E..** 2006. *Dependency Parsing by Inference over High-recall Dependency Predictions*, pages 176-180, New York.

[7] **Carreras, X., Surdeanu, M.,** and **Marquez, L..** 2006. Projective dependency parsing with perceptron, In *Proceedings of CONLL-X,* pages 181-185, New York

[8] **Chang, M.W., Do, Q.,** and **Roth, D..** 2006. A pipeline model for bottom-up dependency parsing. In *Proceedings of CONLL-X*, pages 186-190, New York.

[9] **Cheng, Y., Asahara, M.,** and **Matsumoto, Y..** 2006. Multi-lingual dependency parsing at NAIST. In *Proceedings of CONLL-X,* pages 191-195, New York.

[10] **Corston-Oliver, S.,** and **Aue, A..** 2006. Dependency parsing with reference to Slovene, Spanish and Swedish. In *Proceedings of CONLL-X,* pages 196-200, New York.

[11] **Dreyer, M., Smith, D. A.**, and **Smith, N. A..** 2006. Vine parsing and minimum risk reranking for speed and precision. In *Proceedings of CONLL-X,* pages 201-205, New York.

[12] **Eryiğit, G.,** and **Oflazer, K..** 2006. Statistical dependency parsing of Turkish. In *Proceedings of EACL'06*, pages 89-96, Trento.

[13] **Eryiğit, G., Adalı, E.,** and **Oflazer, K..** 2006a. 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.

[14] **Eryiğit, G., Nivre, J.,** and **Oflazer, K..** 2006b. 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.

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

[16] **Johansson, R.** and **Nugues P..** 2006. Investigating multilingual dependency parsing. In *Proceedings of CONLL-X*, pages 206-210, New York

[17] **Liu, T., Ma, J., Zhu, H.,** and **Li S..** 2006. Dependency parsing based on dynamiz local optimization. In *Proceedings of CONLL-X,* pages 211-215, New York.

[18] **McDonald, R., Lerman, K.,** and **Pereira, F..** 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CONLL-X*, pages 216-220, New York.

[19] **Nivre, J., Hall, J.,Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S.,** and **Marsi, Erwin..** 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering Journal, 13*(1):1-41.

[20] **Oflazer, K., Say, B., Hakkani-Tür D. Z.,** and **Tür, G..** 2003. Building a Turkish treebank. In A. Abeillé, editor, *Treebanks: Building and Using PArsed Corpora*, pages 261-277. Kluwer, London.

[21] **Oflazer, K.,** 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing, 9*(2):137-148.

[22] **Riedel, S., Çakıcı, R.,** and **Meza-Ru,z, I..** 2006. Multi-lingual dependency parsing with incremental integer linear programming. In *Proceedings of CONLL-X*, pages 226-230, New York.

[23] **Say, B..** 2004. Metu-sabancı turkish treebank user guide.

[24] **Schiehlen, M.,** and **Spranger, K..** 2006. Language independent probabilistic context-Free parsing bolstered by machine learning. In *Proceedings of CONLL-X*, pages 231-235, New York.

[25] **Shimizu, N..** 2006. Maximum spanning tree algorithm for non-projective labeled dependency parsing. In *Proceedings of CONLL-X*, pages 241-245, New York.

[26] **Wu, Y.C., Lee, Y.S.,** and **Yang, J.C.**. 2006. The exploriation of deterministic and efficient dependency parsing. In *Proceedings of CONLL-X*, pages 241-245, New York.

[27] **Yüret, D..** 2006. Dependency parsing as a classification problem. In *Proceedings of CONLL-X*, pages 246-250, New York.

[28] **Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel S.,** and **Yüret, D.** 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 915-932. Prague.

[29] **Meral, H. M., Sankur, B., Özsoy, A. S., Güngör, T.,** and **Sevinç, E.** 2009. Natural language watermaking via morphosyntactic alterations. Retrieved from DOI: 10.1016/j.csl.2008.04.001

[30] **Eryiğit, G., İlbay, T.,** and **Can, O. A.** 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages(SPMRL)*, pages 45-55. Dublin, Ireland.

[31] **Eryiğit, G.**. The impact of automatic morphological analysis & disambiguation on dependency parsing of Turkish. 2012.

[32] **Çetinoğlu, Ö.,** and **Kuhn, J.** 2013. Towards joint morphological analysis and dependency parsing of Turkish. In *Proceedings of the Second International Conference on Dependency Linguistics(DepLing)*, pages 23-32. Prague.

[33] **Goenaga, I., Ezeiza, N.,** and **Gojenola, K.** 2013. Exploiting the Contribution of Morphological Information to Parsing: the BASQUE_TEAM system in the SPRML'2013 Shared Task.In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 71-77. Seattle, Washington, USA.

[34] **Durgar El-Kahlout, İ., Akın, A.A.,** and **Yılmaz, E.** 2014. Initial explorations in two-phase Turkish dependency parsing by incorporating constituents. In *First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 82-89. Dublin, Ireland.

[35] **Çetinoğlu, Ö.** Turkish Treebank as a gold standard for morphological disambiguation and its influence on parsing.

[36] **Eryiğit, G**. 2007. ITU Validation Set for Metu-Sabancı Turkish Treebank.