An All-Words Sense Annotated Turkish Corpus

Sinan Akçakaya, Olcay Taner Yıldız
Department of Computer Engineering, Işık University, İstanbul, Turkey
sinan.akcakaya@isik.edu.tr, olcaytaner@isikun.edu.tr

***Abstract*—This paper reports our efforts in constructing of a sense labeled Turkish corpus with respect to Turkish Language Institution's dictionary, using the traditional method of manual tagging. We tagged a pre-built parallel treebank which is translated from the Penn Treebank II corpus. This approach allowed us to generate a full-coverage resource, in which syntactic and semantic information merged. We also provide miscellaneous statistics about the corpus itself as well as its development process.**

***Index Terms*—Natural language processing, Turkish sense annotation, all-words corpus**

I. Introduction

Many natural language processing (NLP) applications use supervised machine learning (ML) algorithms. These methods are data-driven. They require training sets of adequate size to achieve high performance. Thus, the lack of widely available large-scale tagged corpora is the main weakness of supervised techniques. As a result, many of today's NLP systems suffer from data sparseness. This is a fundamental problem, which pervades the discipline of NLP, and is called the knowledge acquisition bottleneck.

Knowledge about word meanings is often required in the domain of NLP. Word Sense Disambiguation (WSD) is a historical task that aims to identify the meaning of words in a given context. WSD systems have potential to make end-to-end language technology applications, such as information retrieval and machine translation, to be more accurate. Unfortunately, until now, WSD has not yet demonstrated real benefits in vivo. This is a consequence of the current performance of WSD, and the main obstacle to the development of higher performance systems is the scarcity of NLP resources such as labeled corpora, part-of-speech (POS) tags, morphological features and syntactic relations. There is an acute need for these kinds of resources, especially for non-English languages.

Based on these requirements, we present a sense tagged Turkish corpus, which has been built on the preceding parallel treebank construction (via alignment from Penn Treebank) and morphological analysis efforts reported in [1] and [2] respectively. In this study, human annotators have disambiguated all words in the input sentences. By virtue of this, the number of words covered is higher than that in the earlier Turkish sense tagged corpora. Another important issue for our dataset is the advantage of ready-to-use root words and other morphological information. It does not require the usually performed preprocessing steps, such as lemmatization, morphological analysis and parsing.

The remaining of the article is organized as follows: First, we give a review of the literature in Section II. Then, we report the former treebank alignment and morphological analysis works performed on the data in Section III. Later, in Section IV, we share our experience on manual tagging of the senses in Turkish Treebank and explain the characteristics of the employed tool. Finally, we conclude and propose subjects for future research in Section V.

II. Related Work

A. Raw corpus

1) English: A raw corpus is the input of a semantic annotation process. As a result of the extensive exploration of the English language in the NLP area, it is not surprising that the most well-known studies are concentrated on English writings. The Brown corpus [3], containing a word balanced selection of present-day American English, totaling about a million words drawn from a wide variety of sources; the Wall Street Journal (WSJ) corpus [4], a library of approximately 30 million words from the WSJ articles; and the British National Corpus [5], a collection of 100 million words extracted from written and spoken English samples from the later part of the 20th century can be listed among the most distinguished English corpora.

2) Turkish: In the domain of Turkish, METU Turkish Corpus [6], a collection of 2 million words of post-1990 written Turkish samples, drawn from several newspapers, magazines and books; the BOUN corpus [7], the largest Turkish web corpus containing around 500 million tokens, formed of major news portals and general Turkish web pages; some other specialized corpora (TS Wikipedia, TweetS, TS Gezi, etc.), published under the TS Corpus project (http://tscorpus.com) are among the widely known available resources.

B. Sense inventory

Sense inventories (e.g. dictionaries, wordnets) are essential for semantic annotation, as they partition the range of the meaning of words into its senses. In contrast to a common dictionary, which provides definitions for words and generally lists them in alphabetical order, in wordnets, all content words (nouns, verbs, adjectives and adverbs) are grouped into sets of cognitive synonyms called synsets. Furthermore, synsets are interlinked by means of conceptual relations (antonym, hypernym, hyponym etc.) and lexical relations. So, wordnets are often considered to be one step ahead of traditional dictionaries and thesauri. After the 2000s, the NLP community widely adopted WordNet [8] instead of traditional dictionaries. Princeton WordNet (PWN) is the most well-known example of

such WordNets, which is a large lexical database of English, created by Princeton University.

Following its success, wordnets for several languages have been created and linked to the original English WordNet. The EuroWordNet [9], providing an interlingual alignment between national wordnets (Dutch, Italian, Spanish, German, French, Czech, and Estonian); the MultiWordNet [10], an Italian wordnet aligned with the Princeton WordNet; BalkaNet [11], wordnets for six European languages (Bulgarian, Czech, Greek, Romanian, Turkish and Serbian) are some of the projects in this direction.

Despite the usefulness of wordnets over the traditional dictionaries, since BalkaNet is not large enough to cover all lexical items in our corpus, we utilized Turkish Language Institution's (a governmental organization, abbreviated as TDK) dictionary. Turkish part of BalkaNet contains only 11,628 synsets with 16,095 literals (members of synsets) in them; whereas TDK dictionary is a collection of 92,371 distinct lemmas (dictionary entry for a word) organized in 121,602 sense entries.

C. Method of annotation

Manual annotation is the most common and reliable way for creating a sense annotated dataset. Unfortunately, however, it is extremely time-consuming. At a throughput of one tagged example per minute [4], depending on the size of the target corpus and the granularity of the sense inventory, generating an annotated corpus can last many years. Allocating workload among multiple annotators relieves the problem, but reveals new issues, such as the competence of annotators and inter-annotator agreement.

To overcome the manual-annotation bottleneck, a number of semiautomatic methods have been proposed. Among them is 'Bootstrapping' [12], which usually starts from a few annotated data and runs a set of one or more classifiers on a vast amount of unlabeled data. As a result of iterative applications of the classifiers, the annotated corpus augments increasingly. Another view in this respect considers Web as a raw corpus and aims to annotate web data with the aid of monosemous relatives [13]. First, one or more search phrases are determined from a dictionary, which uniquely identifies one of the senses of a word. Then, Web is searched for the expressions found in the first step and text snippets are retrieved. Finally, the occurrences of those phrases are replaced with the word in the snippets. The result of this process is an annotated corpus of the word sense that has been focused on.

D. Sense annotated corpora

1) English: NLP community is motivated to build sense tagged corpora, as they have been central to accurate WSD systems. As of today, some annotated corpora are publicly available.

SemCor [14] is the largest and most widely used sense-tagged English corpus, which includes 352 texts marked with around 23,000 different lemmas of 234,000 instances. The texts included in SemCor were drawn from the Brown corpus and a novel, *The Red Badge of Courage*. The word-sense mappings were done with respect to WordNet entries.

The next notable task is the line-hard-serve corpus [15]. It is a typical effort on isolated words. It consists of 4,000 sense-tagged examples of each of the words line (noun), hard (adjective), and serve (verb) with their WordNet senses. Items were extracted from the WSJ, the American Printing House for the Blind, and the San Jose Mercury.

The DSO corpus [16] includes 192,800 occurrences of 191 English words (121 nouns and 70 verbs). This is a noteworthy job, as it focuses on the most frequent words with sufficient numbers of tagged samples. Labeling was done by undergraduates majoring in linguistics and approximately one-man year of effort was spent. Example sentences were picked up from the Brown and WSJ corpora.

Open Mind Word Expert [17] is a corpus of about 90,000 instances of 288 nouns, whose samples were annotated by volunteer web users. The approach relies on the inter-tagger agreement between Web annotators. The authors mention that the system yielded a high quality product at a much lower cost than the traditional method of hiring lexicographers. Its input comes from Penn Treebank, Los Angeles Times, and some other resources.

Additionally, the SENSEVAL competitions (http://www.senseval.org) lead to the periodic release of datasets of high value for the community. It has been held every three years since 1998 to perform a comparative evaluation of WSD systems. Starting with the SENSEVAL-2, several datasets for different languages have been published to perform all-words WSD experiments, together with lexical sample datasets.

2) Turkish: In parallel to increasing concern over Turkish, there have been some efforts towards construction of disambiguated corpora. METU-Sabanci Turkish Treebank [18] is a parsed, morphologically analyzed and disambiguated treebank of 6,930 sentences, which are taken from the METU Turkish Corpus. There are 5,356 different lemmas and 627 of them have 15 or more occurrences. The original treebank does not include sense tags. For the Turkish Lexical Sample Task (TLST) [19] evaluation exercise, which was performed in the SemEval-2007 organization, 5,385 samples of 26 highly ambiguous words from the treebank have been sense tagged based on TDK dictionary.

There is also the Turkish Lexical Sample Dataset (TLSD) [20], which comprises noun and verb sets, each of which has 15 words with high polysemy degree. It has at least 100 samples for each chosen word. The samples have been gathered from various Turkish websites and have been encoded with the senses of TDK by human annotators.

III. Preprocessing steps

Our data have been subjected to the operations listed below, which can be considered as preprocessing steps of our work.

A. Translation

Nowadays, most of the above-mentioned corpora and the other popular ones that are not cited here do not only contain

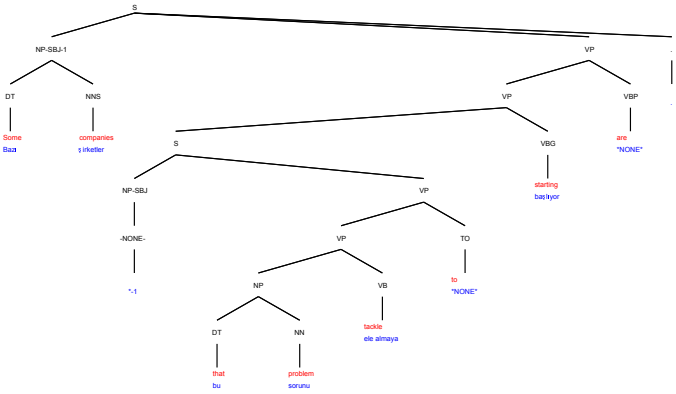

Fig. 1. Syntactic parse of a Penn Treebank sentence "Some companies are starting to tackle that problem" and its Turkish translated form

words and sentences, but have also been employed certain POS tagging and syntactic parsing. Uncovering the syntactic structure of a sentence has clearly many applications in NLP. In fact, many NLP tasks necessitate a syntactic analysis along with a semantic one to obtain worthwhile results. For this reason, we aimed at a full-coverage basis for the Turkish language in which semantic and syntactic understanding are combined. As parsing is a very challenging task, we have exploited a ready-made Turkish treebank [2], which has been aligned from a subset of the original Penn Treebank II corpus [21]. Our converted treebank covers 9,560 sentences with a maximum length of 15 tokens, including punctuation.

The authors report that they converted English parse trees into equivalent Turkish parse trees by applying several transformation heuristics. First, they permuted the subtrees in accordance with the Turkish sentence structure rules. Then, leaf tokens were replaced with the most synonymous Turkish counterparts. As a result, the output of the translation step was not only translated, but also syntactically parsed. Figure 1 illustrate syntactic parse of the Penn Treebank sentence "Some companies are starting to tackle that problem." in the translated form.

B. Morphological analysis

Turkish is an agglutinative language, in which words are formed by attaching derivational and inflectional suffixes to root words. Morphemes added to a word can convert it from a noun to a verb, or vice-versa, or can create adverbs from adjectives. Moreover, during word formation, some letters can disappear or can transform to other characters. Hence, without determining the lemma of a word from its surface form, based on its intended meaning, it is not possible to identify the word correctly and extract candidate senses from a dictionary.

Following the translation, the corpus has been improved with morphological annotations to use in tree-based SMT [2]. In that work, human annotators selected the correct morphological parse from multiple possible analyses returned from the automatic parser. The tag set and morphological representation were quoted from the study, recorded in [18]. Each output of the parser comprises the root of the word, its part-of-speech tag and a set of morphemes, each separated with a + sign.

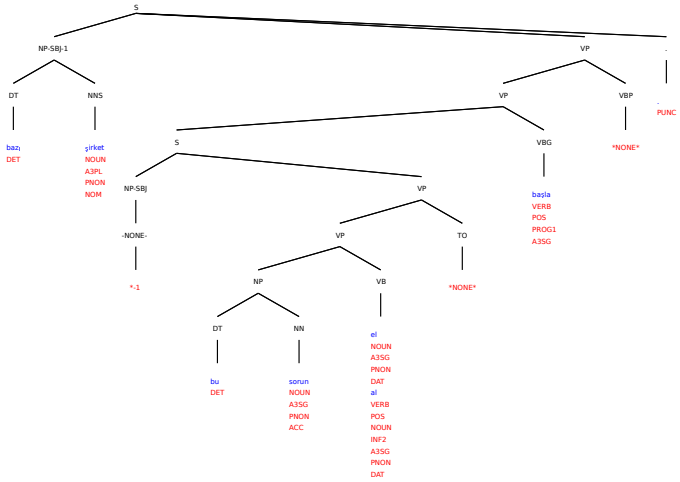

Fig. 2. Morphologically disambiguated form of Turkish sentence in Figure 1

Figure 2 illustrates the morphologically disambiguated form of the translated Penn Treebank sentence given in Figure 1.

IV. Corpus construction

In our study, the overall process has been carried out by human annotators in a manual style. For this task, experts have been assisted by a custom application, which is designed for manual sense annotation.

A. Basic data format

Through the preliminary phases (translation and morphological analysis) and the disambiguation work explained here, common materials were treated by remaining highly faithful to the standard Penn Treebank notation of syntactic bracketing in the backend. Each of these successive studies enriched the documents in terms of its application goals, based on the previous ones. We extend the original format with the relevant information, given between curly braces. For example, the word "problem" in the sentence shown in Figure 1 in the standard Penn Treebank notation, may be represented in the data format provided below:

```
(NN problem)
```

After all levels of processing are finished, the data structure stored for the same word has the following form in our system.

```
(NN {turkish=sorunu} {english=problem}
{morphAnalysis=sorun+NOUN+A3SG+PNON+ACC}
{metaMorphemes=sorun+yH}
{semantics=TUR10-0703650})
```

B. Sense Inventory

We stored the TDK dictionary in XML format that is fairly similar to the BalkaNet's. In our format, units that constitute the vocabulary are possible meanings of the words. We named these units 'synsets', as conventional in the domain of wordnets, but our synsets are not merged or interlinked with other synsets. In fact, we have not made extra processing on the original dictionary; instead, we transfigured into a shape

which resembles the BalkaNet format. The structure of a sample synset is as follows:

```
<SYNSET>
<ID>TUR10-0066140</ID>
<LITERAL>baba
<SENSE>1</SENSE>
</LITERAL>
<POS>n</POS>
<DEF>...</DEF>
<EXAMPLE>...</EXAMPLE>
</SYNSET>
```

Each entry in the dictionary is enclosed by <SYNSET> and </SYNSET> tags. Synset members are represented as literals and their sense numbers. In contrast to BalkaNet, where synonym literals are joined within a synset, our synsets can have only one literal. <ID> shows unique identifier given to the synset. <POS> and <DEF> tags denote part of speech and definition, respectively. As for the <EXAMPLE> tag, it gives a sample sentence for the synset.

C. Annotation tool

The annotators use a custom application (written in Java) for browsing sentences and annotating them with senses. Our application has been integrated as a part of the Işık University NLP Toolkit, a collection of in-house tools designed for NLP studies. The toolkit is freely available and can be downloaded from the web page[1]. Thus, we could use the same infrastructure with the prior treebank alignment and morphological analysis tasks.

The current implementation of the application is designed for the import of text files that adhere to our Penn Treebank style data format (that is, translated and morphologically analysed). Once a pre-processed sentence has been imported into the semantic editor, the human annotator is presented with the visualized syntactic parse tree of that sentence. Annotators can click on leaf nodes (words), but they are not allowed to make any changes such as rotating or deleting nodes. When a word is selected, a drop-down list is displayed, in which all available TDK entries of the selected lemma are listed. In all-words annotation tasks, list of words to be tagged, and therefore also the candidate senses, are unpredictable. Our application handles sense extraction on behalf of the annotators.

Each sense result shown to the annotators is populated with its POS and a sample sentence (which are already available in the TDK dictionary). This becomes a considerable aid for the annotators in deciding which sense to assign to a target word. Moreover, sense options whose POS do not agree with the word's POS, are disabled (are shown but not selectable) to facilitate the task. Just after the selection of the most appropriate sense, the drop-down list is hidden and the ID of the submitted synset is displayed under the word. Figure 3 shows a screenshot from the system interface, depicting the screen presented to the raters when tagging the noun "sorun".

[1]http://haydut.isikun.edu.tr/nlptoolkit.html

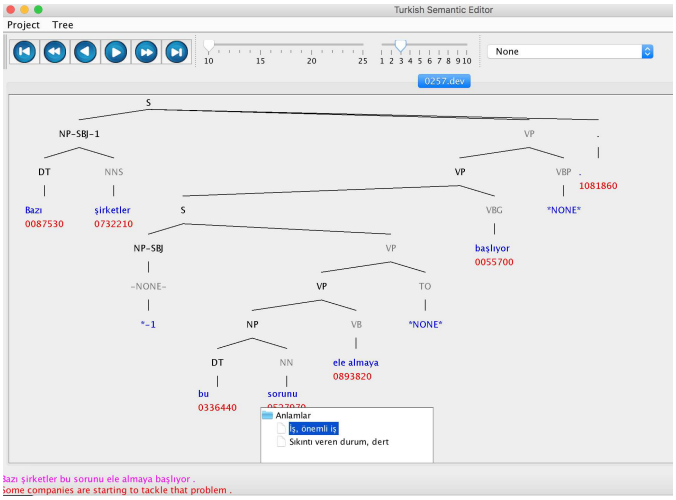

Fig. 3. A screenshot from the system interface

D. Extracting candidate senses from the dictionary

1) Simple candidate senses: The available senses of a word are pulled by querying its lemma in the lexicon. For example, in the converted sentence shown in Figure 2, the Turkish noun "sorunu" can be morphologically decomposed in three different ways, illustrated below.

sorun + NOUN + A3SG + PNON + ACC (the problem)
sorun + NOUN + A3SG + P3SG + NOM (her/his problem)
soru + NOUN + A3SG + P2SG + ACC (your question)

As mentioned before, morphological disambiguation has been done by human annotators in the past study reported in [2]. In the course of annotation, our system queries the dictionary with either "soru" (question) or "sorun" (problem) according to the selected morphological analysis. Verbs are different from the other word types, as their dictionary forms are found by appending -mek or -mak suffixes to the root.

Due to the agglutinative nature of Turkish, in some cases, the output of the morphological parsers does not cover the correct dictionary form. With an example sentence given below, we demonstrate this issue and the solution:

Şimdilik, her iki şirket özel planları hakkında sessizliklerini koruyor

(For now, both companies are keeping quiet about their specific plans)

TABLE I
MORPHOLOGICAL ANALYSIS OF THE TARGET WORD "SESSIZLIKLERINI" AND CONSTRUCTED SURFACEFORMS AFTER APPENDING EACH METAMORPHEME

ses+NOUN+A3SG+PNON+NOM	ses
^DB+ADJ+WITHOUT	sessiz
^DB+NOUN+NESS	sessizlik
+A3PL+P3PL	sessizlikleri
+ACC	sessizliklerini

The root word "ses" (voice) obtained by the automatic parser is the basis of new words that can be formed by adding combinations of suffixes to them. However, for the word "sessizliklerini", the correct dictionary form is "sessizlik" (quite, silence), which is a stand-alone word on its own. This

situation occurs for the words that have been formed by adding derivational suffixes to a base word. To handle this problem, the annotation tool questions the dictionary with all possible lemmas that can be derived by appending metamorphemes to the root word (Table I). Following the solution presented here, the application pulls the senses of the root word "ses" as well as each of the derived items given below:

sessiz (silent), sessizlik (quite, silence), sessizlikleri and sessizliklerini

A candidate sense list is the union of senses returned by each query.

2) Senses of collocations: Another issue that must be handled by the tool is collocations. Many English words have a multi-word translation into Turkish and they need special attention to obtain a sense list. In this case, we extract candidate senses for each constituent, as explained above, and merge them in a list. This list still misses the overall meaning of the multi-word expression. As a solution, we take cartesian product of derived forms of each singleton and poll the dictionary for each combination. If any senses are found, we add them into the list. For instance, consider the following parallel sentences:

İki yüzlü ile dostluk kurma

(Do not make friends with a hypocrite)

In this sentence, the collocation is "iki yüzlü". It corresponds to "hypocrite" in the English side. Potential individual words are:

iki (two)

yüz (face), yüzlü (faced)

There are two possible cartesian product results, given below:

iki yüz (is not a Turkish collocation)

iki yüzlü (a Turkish collocation)

Only "iki yüzlü" is a real Turkish collocation among them. Eventually, available senses for the item "iki yüzlü" is composed of the senses of the simplexes "iki", "yüz" and "yüzlü", plus returned for multi-word "iki yüzlü".

E. Some features and statistics about the corpus

In producing the corpus, commentators are asked to select the most appropriate sense for each word in a given sentence. Unambiguous entities, such as proper nouns, times, dates, punctuations and numbers, are also included in the task. They are assigned with the specific synsets reserved for each of them. As a consequence, word frequencies and the coverage of senses are not balanced. The upshot of the one year effort is a corpus of about 83,500 word occurrences, of which 51,117 are polysemous. There are 7,595 distinct content words and 202 of them occur 50 or more times in the corpus. The average number of samples per lemma is equal to 7.8. Table II lists the distribution of words and distinct lemmas by their grammatical categories.

In the dictionary of TDK, the range of the number of a word's possible senses is quite irregular. While many of them are monosemous, some of them may have up to 50 senses. Hereby, the time required to annotate one target word diverges from a few seconds (for monosemous and homonyms) to several minutes (for subtle cases where even disagreements arise among annotators). We list the 10 most-frequent words with their rankings and number of candidate senses in Table III. Also, several features of the resulting corpus are given in Table IV in comparison with the other Turkish sense tagged corpora.

TABLE II
DISTRIBUTION OF WORD TYPES

Word Type	Sample Size	Distinct Sample Size
Noun	33,320	5,882
Verb	11,981	747
Adjective	9,591	739
Adverb	3,704	227
Number, Range	4,460	
Punctuation	14,372	
Pronoun, etc.	6,046	
Total	83,474	7,595

TABLE III
10 MOST-FREQUENT WORDS AND THEIR RANKINGS

Word	# of Occurence	# of Senses
olmak (be, happen)	1,072	25
etmek (auxiliary verb)	765	9
dolar (dollar)	622	1
hisse (share, stock)	616	3
bay (mr.)	481	3
şirket (company, firm)	391	1
satmak (sell)	383	5
milyon (million)	380	2
yapmak (do, make)	375	20
demek (say, tell)	339	15

TABLE IV
COMPARISON OF TURKISH SENSE ANNOTATED CORPORA

Corpus	Example	Lemma	Coverage	Syntactic Parse
Semeval-2007	5,385	26	Lexical sample	Available
TLSD	3,616	35	Lexical sample	Unavailable
Our corpus	83,474	7,524	All-words	Available

Annotation process has been performed by five annotators majoring in computer engineering. At the beginning of the task, workload was distributed among them. Each annotator has engaged in labeling distinct set of sentences. This means that each item in the corpus has been labeled once. After the entire corpus has been tagged, annotators have been asked to label the same 250 sentences from the corpus to evaluate the quality of the annotation task performed so far. We have measured 82% inter-annotator agreement on 2,054 words, which are organized in the common test set. The items, for which disagreements have arisen through evaluation, have been assigned the sense selected by the majority of the five human annotators.

In the scope of this work, we have also made WSD baseline calculations on this Turkish dataset. With the employment of 10-fold cross-validation runs through calculations, we have obtained 25.04% accuracy for the random baseline and 61.26% for the first sense baseline. For the above-mentioned Turkish corpus TLSD [20], authors reported 29.15% accuracy for the most frequent sense (MFS) baseline, which is fairly low compared to 61.26%. This is originated from the equally

balanced structure of TLSD. In the extreme case of exactly equal frequencies of the senses of words, the random baseline performs with an accuracy equal to the MFS baseline. Note that 29.15% is close to our random baseline 25.04%. However, 61.26% is comparable with the MFS baseline values of 57% and 60.9%, calculated on English all-words corpora used in SENSEVAL-2 and SENSEVAL-3 competitions respectively.

V. Conclusion and future work

In this paper, we reported our experience on manual tagging of TDK senses in a Turkish-English parallel treebank, aligned from Penn Treebank. This corpus had already been parsed and enhanced with morphological features before the semantic annotation process presented here. From the initial translation endeavors to the current sense labeling degree, common text files have been processed in a progressive manner. That is, each distinct study made full use of the previous one without corrupting it and produced a cumulative data set. Such a technique of exploiting resources at hand, either a treebank in a different language or morphological analyses in native language, proved to be a useful method that accelerates the corpus construction process.

During the process, same software infrastructure have been used with the previous tasks. By means of this, users could make necessary corrections when they realized errors in the former stages (e.g., simple misspellings in the translation, mistakes in the morphological disambiguation, etc.) by shuttling through the phases. Loop style of dealing with data like this, verified the accuracy and consistency of the overall process.

Turkish is one of the languages that need much more linguistic resources to speed up NLP research. Creation of this dataset will contribute to this call, offering an all-words sense annotated corpus. We hope that this corpus will also be a useful resource for various NLP studies, since it is a full-coverage material that provides syntactic parse and morphological analysis together with sense annotations. Such an all-words Turkish corpus has not been recorded yet. Therefore, it has significance for NLP studies concentrating on Turkish.

The resulting corpus has advantages over the other annotated corpora of Turkish in several respects: (1) Word coverage is much more extensive, since all words are labeled. This enables us to make WSD experiments in an all-words approach. (2) Root forms and morphological structure are on hand. This eliminates the need for an external morphological analyser and disambiguator. (3) Syntactic features available in the parsed sentences makes it possible to acquire more information about the words. Hence, we can combine semantic and syntactic features in a variety of NLP applications. As a future work, we plan to expand the size of the corpus by following the same manner of work and perform WSD experiments on it, with various classifiers and feature sets.

Acknowledgements

This work was supported by Tübitak project 116E104 and Işık University BAP project 15B201.

References

[1] O. T. Yildiz, E. Solak, O. Gorgun, and R. Ehsani, "Constructing a Turkish-English parallel treebank," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 112–117.

[2] O. Gorgun, O. T. Yildiz, E. Solak, and R. Ehsani, "English-Turkish parallel treebank with morphological annotations and its use in tree-based smt," in *International Conference on Pattern Recognition and Methods*, Rome, Italy, 2016, pp. 510–516.

[3] H. Kucera and W. N. Francis, "Computational analysis of present-day american english," *International Journal of American Linguistics*, vol. 35, pp. 71–75, 1969.

[4] E. Charniak, D. Blaheta, N. Ge, K. Hall, J. Hale, and M. Johnson, "Bllip 1987-89 wsj corpus release 1," Philadelphia, Tech. Rep., 2000.

[5] J. Clear, "The british national corpus," in *The Digital Word: Text-Based Computing in the Humanities*. Cambridge, MA, USA: MIT Press, 1993, pp. 163–187.

[6] B. Say, D. Zeyrek, K. Oflazer, and U. Özge, "Development of a corpus and a treebank for present-day written turkish," in *Proceedings of the Eleventh International Conference of Turkish Linguistics*, Eastern Mediterranean University, Cyprus, August 2002, pp. 183–192.

[7] H. Sak, T. Güngör, and M. Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *Proceedings of the 6th international conference on Advances in Natural Language Processing*, Gothenburg, Sweden, 2008, pp. 417–427.

[8] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Wordnet: An online lexical database," *International Journal of Lexicography*, vol. 3, pp. 235–244, 1990.

[9] P. Vossen, *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1998.

[10] E. Pianta, L. Bentivogli, and C. Girardi, "Multiwordnet: Developing an aligned multilingual database," in *Proceedings of the 1st International Conference on Global WordNet*, Mysore, India, 2002, pp. 21–25.

[11] S. Stamou, K. Oflazer, K. Pala, D. Christodoulakis, D. Cristea, D. Tufis, S. Koeva, S. Totkov, D. Dutoit, and M. Grigoriadou, "Balkanet: A multilingual semantic network for balkan languages," in *Proceedings of the First International WordNet Conference*, Mysore, India, 2002, pp. 21–25.

[12] R. Mihalcea, "Co-training and self-training for word sense disambiguation," in *Proceedings of the 8th Conference on Computational Natural Language Learning*, Boston, MA, USA, 2004, pp. 33–40.

[13] ——, "Bootstrapping large sense tagged corpora," in *Proceedings of the 3rd International Conference on Languages Resources and Evaluations*, Spain, 2002, pp. 33–40.

[14] G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker, "A semantic concordance," in *Proceedings of the ARPA Workshop on Human Language Technology*, Stroudsburg, PA, USA, 1993, pp. 303–308.

[15] C. Leacock, G. Towell, and E. Voorhees, "Corpus-based statistical sense resolution," in *Proceedings of the ARPA Workshop on Human Language Technology*, Princeton, NJ, 1993, pp. 260–265.

[16] H. T. Ng and H. B. Lee, "Integrating multiple knowledge sources to disambiguate word senses: An examplar-based approach," in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, USA, 1996, pp. 40–47.

[17] T. Chklovski and R. Mihalcea, "Building a sense tagged corpus with open mind word expert," in *Proceedings of ACL 2002 Workshop on WSD: Recent Successes and Future Directions*, Philadelphia, PA, 2002, pp. 116–122.

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

[19] Z. Orhan, E. Çelik, and N. Demirgüç, "Turkish lexical sample task," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, 2007, pp. 59–63.

[20] B. İlgen, E. Adalı, and A. C. Tantuğ, "Building up lexical sample dataset for turkish word sense disambiguation," in *IEEE International Symposium on Innovations in Intelligent Systems and Applications*, Trabzon, Turkey, July 2012, pp. 1–5.

[21] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of english: The penn treebank," *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.