English-Turkish Parallel Semantic Annotation of Penn-Treebank

Bilge Nas Arıcan
Starlang Yazılım Danışmanlık, Turkey
bnarican@gmail.com

Özge Bakay
Boğaziçi University, Turkey
ozge.bakay@boun.edu.tr

Begüm Avar
Boğaziçi University, Turkey
begum.avar@boun.edu.tr

Olcay Taner Yıldız
Işık University, Turkey
olcaytaner@isikun.edu.tr

Özlem Ergelen
Boğaziçi University, Turkey
ozlem.ergelen@boun.edu.tr

Abstract

This paper reports our efforts in constructing a sense-labeled English-Turkish parallel corpus using the traditional method of manual tagging. We tagged a pre-built parallel treebank which was translated from the Penn Treebank corpus. This approach allowed us to generate a resource combining syntactic and semantic information. We provide statistics about the corpus itself as well as information regarding its development process.

1 Introduction

Parallel corpora, which are a collection of texts in one language and their translations in at least one other, can be used in a variety of fields, such as translation studies and contrastive linguistics. They are used for many different purposes including creating new linguistic resources such as lexicons and WordNet (Petrolito and Bond, 2014). As for the relationship between parallel corpora and natural language processing (NLP) studies, in addition to the fact NLP studies use parallel corpora as material bases or testing arenas, NLP studies also contribute to the development of corpora in many areas, especially in corpus annotation.

In this paper, we present a sense-tagged English-Turkish parallel corpus, which is the only corpus for the English-Turkish combination having both semantic and syntactic information. It has been built on the preceding parallel treebank construction and morphological analysis efforts reported in (Yildiz et al., 2014) and (Gorgun et al., 2016). The aim of this study is to investigate the possibility of a parallel semantic annotation for an English-Turkish corpus. The motivation behind the study is the potential contribution of this parallel semantic annotation to several NLP tasks such as automatic annotation, statistical machine translation and word sense disambiguation.

This paper is organized as follows: We give background information about lexical semantics in Section 2 and present the related work in Section 3. The details of our corpus and how it is constructed are given in Section 4. We provide the annotation statistics about the corpus in Section 5 and conclude in Section 6.

2 Lexical Semantics

In linguistics, lexical semantics is the study of word meaning. The main challenge in this field is generated from 'polysemy', which is the term used for the phenomenon of a single orthographic word having multiple, interrelated senses. In classical dictionaries, these senses are listed under a single lexical entry and, as stated in (Firth, 1957), "You shall know a word by the company it keeps", that is, only with the help of the context one can pin down the particular sense in which a word is used. A further challenge in the field stems from collations, i.e. groups of words having "a unitary meaning which does not correspond to the compositional meaning of their parts" (Saeed, 1997).

Hence, as far as compositionality is considered to be crucial to semantic analysis, there are two central concerns for the semanticist: (i) At the lexical level, choosing the correct sense of a given word within a context, and (ii) at the sentence level, determining how a particular combination of words should be interpreted.

Languages also differ in terms of how lexical items are combined, which is directly related to how compositionality is to be interpreted. Therefore, the success and adequacy of a multi-

lingual semantic analysis not only requires taking "into consideration the multitude of different senses of words across languages", but also "effective mechanisms that allow for the linking of extended word senses in diverging polysemy patterns" (Boas, 2005).

When it comes to interlingual semantics studies, even further complications arise. For one, there is a huge discrepancy between languages in terms of which semantic components they lexicalize. For instance, in analytic languages like English, functional morphemes are free forms, such as determiners and appositions, whereas in agglutinative languages, such as Turkish, syntactic relations are expressed mainly via affixation. Hence, a single orthographic word in Turkish may correspond to a phrase consisting of a combination of multiple free morphemes in English.

3 Related Work

In this section, we present previous work and provide a comparison of our corpus with other corpora mainly with reference to their sense annotation process and the number of annotated words.

3.1 English Semantically-Annotated Corpora

Among many corpora concentrated on English is SemCor (Miller et al., 1993), which is the most widely-used and largest sense-tagged English corpus with 192,639 instances. SemCor's input comes from the novel of The Red Badge of Courage and the Brown corpus, which presents one million words in contemporary American English obtained from various sources. As for the word-sense mappings, they were done based on WordNet entries.

Another significant study in this area is the line-hard-serve corpus (Leacock et al., 1993). Having extracted its data from three different resources, it is comprised of 4,000 sense-tagged examples of each of the words line (noun), hard (adjective), and serve (verb), which are also mapped with their WordNet senses.

Table 1 shows the English partition of our corpus in comparison with the other English sense-tagged corpora. Our English corpus can be considered as a noteworthy example in terms of its target, the number of annotated words and the version of WordNet used. Having all words annotated by using the latest version of WordNet (WN 3.1), our corpus annotates 41,986 words in total.

3.2 Multilingual Semantically-Annotated Corpora

Among interlingual studies aligned with SemCor, there is the English/Italian parallel corpus called MultiSemCor (Bentivogli et al., 2005), which is aligned at the word level and annotated with PoS, lemma and word sense. Their corpus contains around 120,000 English words annotated, approximately 93,000 of which are transferred to Italian and annotated with Italian word senses. Another important project is by (Lupu et al., 2005). Targeting all words to be annotated, their corpus, SemCor-En/Ro, contains around 48,000 tagged words in Romanian.

The comparison of our multilingual corpus with the other multilingual sense-tagged corpora is given in Table 2. Our corpus is notable when compared to the other corpora for three main reasons; first, it uses the latest version of WordNet (WN 3.1) unlike many other multilingual corpora; second, the total number of words annotated for both languages in our corpus is substantial for a preliminary work; third, it is the first parallel semantically annotated corpus for English-Turkish language pair.

3.3 Turkish Semantically-Annotated Corpora

METU-Sabanci Turkish Treebank (Oflazer et al., 2003), which is a parsed, morphologically-analyzed and disambiguated treebank of 6,930 sentences, is a substantial corpus for Turkish. The sentences were extracted from the METU Turkish corpus, which is a compilation of 2 million words from written Turkish samples gathered from several resources (Say et al., 2002). In these sentences, 5,356 lemmas are annotated, with 627 of them having at least 15 occurences.

Another exemplary corpus for Turkish is the Turkish Lexical Sample Dataset (TLSD) (İlgen et al., 2012). It includes noun and verb sets and both sets have 15 words each with high polysemy degree. An important strength of this corpus is that each word has at least 100 samples which were gathered from various Turkish websites and encoded with the senses of TDK (the Turkish Language Institution's dictionary) by human interpreters.

Our Turkish corpus, on the other hand, is prominent among the current Turkish corpora. As Table 3 suggests, it is the only Turkish corpus both an-

Table 1: Comparison of English sense-annotated corpora

Corpus	# Words Tagged	WordNet	Target
SemCor3.0-all (Miller et al., 1993)	192,639	WN 3.0	all
SemCor3.0-verbs (Miller et al., 1993)	41,497	WN 3.0	verbs
Gloss Corpus (Miller et al., 1993)	449,355	WN 3.0	some
Line-hard-serve (Leacock et al., 1993)	4,000	WN 1.5	some
DSO corpus (Ng and Lee, 1996)	192,800	WN 1.5	nouns, verbs
Senseval 3 (Snyder and Palmer, 2005)	2,212	WN 1.7.1	all
MASC (Ide, 2012)	100,000	WN 3.0	verbs
SemEval-2013 Task 13 (Jurgens and Klapaftis, 2013)	5,000	WN 3.1	nouns
Our corpus	41,986	WN 3.1	all

Table 2: Comparison of multilingual sense-annotated corpora

Corpus	# Words Tagged	Languages	WordNet	Target
MultiSemCor	92,420	Italian	MultiWN	all
(Bentivogli et al., 2005)	119,802	English	WN 1.6	
SemCor-En/Ro	48,392	Romanian	BalkaNet	all
(Lupu et al., 2005)	n/a	English	WN 2.0	
NTU-MC	36,173; 27,796	Chinese; Indonesian	COW; WN Bahasa	all
(Tan and Bond, 2012)	15,395; 51,147	Japanese; English	Jpn WN; PWN	
SemEval-2013 Task 12	3,000; 3,000	French; Spanish	BabelNet	all
(Navigli et al., 2013)	3,000; 4,000	German; Italian		
Our corpus	61,127; 41,986	Turkish; English	KeNet 1.0; WN 3.1	all

Table 3: Comparison of Turkish sense-annotated corpora

Corpus	# Words Tagged	# Lemma	Target	Syntactic Parse
SemEval-2007 (Orhan et al., 2007)	5,385	26	noun; verbs	Available
TLSD (İlgen et al., 2012)	3,616	35	noun; verbs	Unavailable
Our corpus	61,127	7,017	all	Available

notating all words and providing their syntactic information and it annotates by far the largest number of words in total, 61,127. Second, it is also the only Turkish corpus which is parallel annotated.

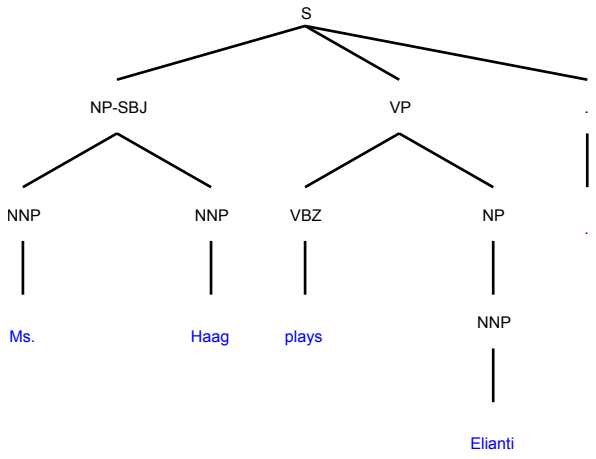
4 Corpus

In this section, we describe how the data in our corpus were extracted and organized, give details of our annotation tool, explain how the data in both Turkish and English partitions were annotated, give an account of our data format, and finally, evaluate our annotation.

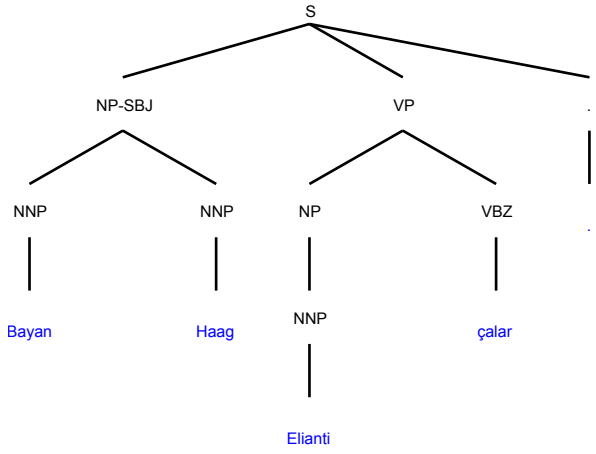
4.1 Preliminary Corpus

As a preliminary work for our corpus, we disambiguated the Turkish-English parallel Treebank (Yildiz et al., 2014) where the English parse trees were converted into their equivalent Turkish parse trees with the application of several transformation heuristics. First, the subtrees were permuted with reference to the Turkish sentence structure rules. Then, leaf tokens were replaced with the most synonymous Turkish counterparts. Finally, an output which was both translated and syntactically-parsed was formed.

Regarding the differences related to syntax, one should note that the majority of Turkish sentences have the Subject-Object-Verb word order whereas most English sentences have Subject-Verb-Object order. When translating English trees, they permute its subtrees to reflect the change of constituent order in Turkish. For example, when translating the sentence in Figure 1(a), VBZ and NP subtrees are exchanged so that the correct con-



(a)



(b)

Figure 1: An example English sentence from Penn-Treebank corpus (a) and its translated form (b)

stituent order in Turkish is constructed in the translated form (Figure 1(b)).

They also use the *NONE* tag when they cannot use any direct gloss for an English token. The semantic aspects expressed by prepositions, modals, particles and verb tenses in English in general correspond to specific morphemes attached to the corresponding word stem in Turkish. By using *NONE* tag, permuting the nodes and choosing the full inflected forms of the glosses in the Turkish tree, they have a working method to convert subtrees to an inflected word.

Following the translation phase, the corpus has been improved with morphological annotations to use in tree-based statistical machine translation (Gorgun et al., 2016). In that work, human annotators selected the correct morphological parse from multiple possible analyses returned from the

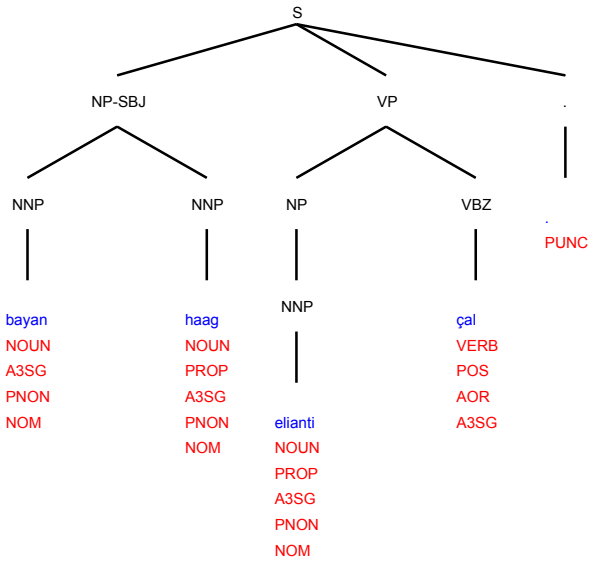


Figure 2: Morphologically-disambiguated form of the sentence in Figure 1(a)

automatic parser. The tag set and morphological representation were quoted from the study reported in (Oflazer et al., 2003). Each output of the parser comprises the root of the word, its part-of-speech tag and a set of its morphemes, each separated with a "+" sign. Figure 2 illustrates the morphologically disambiguated form of the sentence in Figure 1(a).

4.2 Annotation Tool

The annotators use a custom application (written in Java) for browsing sentences and annotating them with senses. The toolkit is freely available [1]. The current implementation of the application is designed for the import of text files that adhere to the Penn Treebank data format (that is, translated and morphologically analyzed).

Once a pre-processed sentence has been imported into the semantic editor, human annotators are presented with the visualized syntactic parse tree of that sentence. Annotators can click on leaf nodes, which correspond to the words. When a word is selected, a drop-down list is displayed, in which all the available WordNet entries of the selected lemma are listed. Figure 3 shows a screenshot from the system interface, depicting the screen presented to the annotators when annotating the verb "çalar" in the Turkish sentence "Bayan Haag Elianti çalar." Right after the selection of the most appropriate sense, the drop-down

[1] https://github.com/olcaytaner/DataCollector

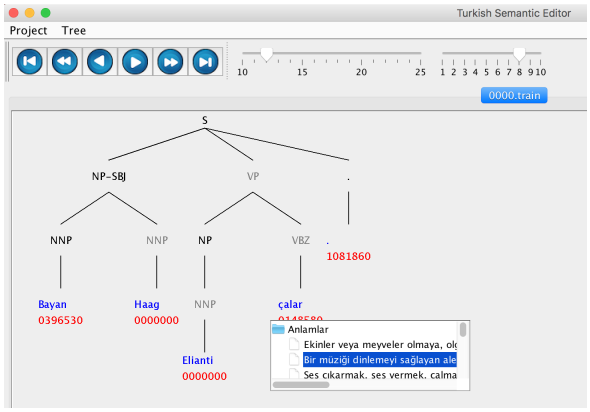


Figure 3: A screenshot from the system interface

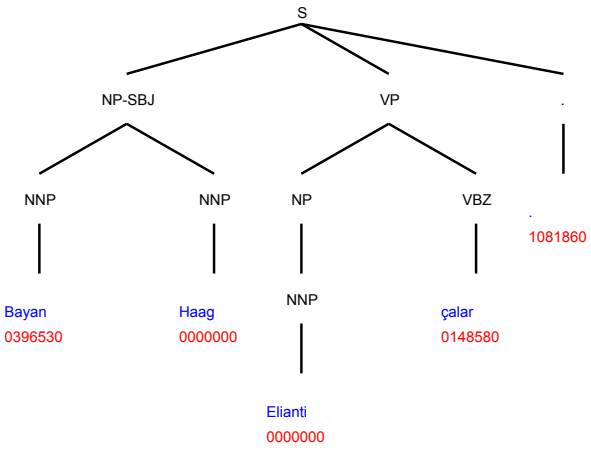


Figure 4: Sense-annotated form of the Turkish sentence in Figure 1(a)

list is hidden and the ID of the submitted synset is displayed under the word. Figure 4 shows the sense-annotated form of the Turkish sentence in Figure 1(a).

4.3 Turkish Sense Annotation

4.3.1 Sense Inventory

For the Turkish sense annotation, the Turkish WordNet KeNet 1.0 (Ehsani et al., 2018) was used. KeNet was stored in an XML format that is quite similar to BalkaNet's (Stamou et al., 2002). The structure of a sample synset is as follows:

```
<SYNSET>
<ID>0066140</ID>
<SYNONYM>
<LITERAL>baba<SENSE>1</SENSE>
</LITERAL>
<LITERAL>peder<SENSE>1</SENSE>
</LITERAL>
```

Table 4: Unambiguous entities in the Turkish WordNet

Id	Entity
0000000	Proper noun
0000003	Time
0000004	Date
0000006	Hash tag
0000007	E-mail
0000010	Integer
0000011	Ordinal number
0000013	Percentage
0000015	Rational number
0000018	Interval
0000020	Real number

```
<SYNONYM>
<POS>n</POS>
<DEF>Çocuğu olmuş erkek</DEF>
<EXAMPLE>Babasını çok sever.
</EXAMPLE>
</SYNSET>
```

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

For the Turkish side of the corpus, unambiguous entities, such as proper nouns, numbers or dates, are also included in the task where they are assigned with the IDs for their specific synsets (See Table 4). For instance, in Figure 4, the words "Bayan" and "Elianti" are assigned the ID of "0000000", which is the synset ID for proper nouns.

4.3.2 Extracting Candidate Sense List

The available senses of a word are obtained by querying its root word in KeNet. For example, in the converted sentence shown in Figure 2, the Turkish verb "çalar" can be morphologically decomposed in three different ways as illustrated below.

çal + VERB + POS + AOR + A3SG (plays)
çal + VERB + POS + AOR∧DB + ADJ + ZERO (playing X)
çalar + NOUN + A3SG + PNON + NOM (player)

As mentioned before, morphological disambiguation has been done by human annotators in the past study reported in (Gorgun et al., 2016). In the course of annotation, our system queries the dictionary with "çal" (play) or "çalar" (player) according to the selected morphological analysis. This morphological disambiguation prior to the annotation process is crucial especially in agglutinative languages such as Turkish. Thanks to this morphological disambiguation, the annotation process has been accelerated since the annotators have been provided with shorter lists of possible senses depending on the part of speech (POS) of the word being annotated in the given sentence. For example, when the annotator is to annotate the word "çalar" (play) in Figure 4, the software lists its senses as a verb and excludes the other senses provided by other POSs such as the sense(s) of "çalar" (player) as a noun.

Another issue that must be handled by the sense disambiguation tool is collocations. Many English words have a multi-word translation into Turkish and they need special attention to obtain a sense list. As a solution, we take cartesian product of derived forms of each word and search the WordNet for each combination. If any sense is found, we add it into the sense lists of the words that are included in the collocation. For instance, consider the following parallel sentences:

Hisse senedini sattı.

He sold the stock.

In the Turkish sentence, there is one collocation, namely "hisse senedi" which corresponds to "stock" in the English partition. After taking all the possible productions of the two words, "hisse" and "senedini" ("hisse senet", "hisse senedi", "hisse senedini"), the available senses displayed in the droplist for the word "hisse" contain both the possible senses of the simplex "hisse" and the ones corresponding to the collation of "hisse senedi".

4.4 English Sense Annotation

4.4.1 Sense Inventory

For the English sense annotation, we use Princeton WordNet (PWN) version 3.1. Although PWN does not provide a web page for obtaining synsets and/or their relations, the data files are present. After retrieving the synset data files from the site, we constructed a WordNet XML file similar to the Turkish one as given in Section 4.3.1:

```
<SYNSET>
<ID>10100638</ID>
<LITERAL>father<SENSE>1</SENSE>
</LITERAL>
<LITERAL>begetter<SENSE>1</SENSE>
</LITERAL>
<POS>n</POS>
<DEF>a male parent</DEF>
<EXAMPLE>...</EXAMPLE>
</SYNSET>
```

4.4.2 Extracting Candidate Sense List

For the English partition, extracting simple senses is much easier. We only ask for the available senses of the English word in PWN. Complexities arise for verbs marked for third person (-s), gerund (-ing), past participle (-ed); and for adjectives in comparative (-er) or superlative (-est) forms. For those cases, we strip down the affixes and search for the root form in PWN. For irregular forms (such as irregular verbs), we use the exception list of PWN to get the root forms.

Whereas function words are left unannotated, their copular or lexical counterparts are annotated. For instance, while the auxiliary verbs "be" and "have" are not annotated with a sense, their copular or lexical counterparts, such as "have" in the example of "The company had a loss", have been assigned a sense by the annotators. 868 of all the occurrences of "be" and "have" are lexical; and thus, were annotated with a sense.

For collocations, the situation is again easy for the English partition. We search for 2 or 3 word collocations in PWN with respect to the adjacent words of the current word. For instance, consider the sentence "They get up early". While showing the sense list of "get", we do not only show the sense list of "get" in isolation, but also add the senses of "get up" to that list. There are also collocations written with a hyphen in-between. For the ones listed as a single entry in the dictionary, such as "way-out", we add the senses under each word included in the collocation. The number of that kind of collocations with senses annotated is 219. However, the ones that cannot be treated as single lexical items, such as "three-months", were left unannotated. In total, 998 collocations with a hyphen could not be assigned a sense.

For the sake of consistency, since the corpus has a number of recurring words, annotators have compiled a list of the most frequently occurring 82 polysemous words, with multiple sense definitions

differing only slightly from each other. They have then decided on what sense is to be chosen and assigned to these words, and in which contexts. In addition, the annotators have agreed on certain conventions in annotating quantificational expressions, including numerals. The preparation of such a convention-guide, which is used as a sense-annotation-lexicon, helped each annotator to consistently select the same sense for a given word occurring in the same context and increased the inter-annotator agreement rate.

4.5 Data Format

In order to be able to process further, we remain highly faithful to the standard Penn Treebank notation of syntactic bracketing in the backend. We extend the original format with the relevant information, given between curly braces. For example, the word "plays" in the sentence shown in Figure 1 in the standard Penn Treebank notation, may be represented in the data format provided below:

```
(VBZ plays)
```

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

```
(VBZ {turkish=çalar}{english=plays}
{turkishSemantics=0703650}
{englishSemantics=15161405-n})
```

If there are multiple words on the Turkish side, the senses of each word is separated via a dollar sign:

```
(JJR {turkish=daha nazik}
{english=gentler}
{turkishSemantics=0178860$0572140}
{englishSemantics=01458191-s})
```

except collocations, for which a single sense ID is sufficient:

```
(NN {turkish=hisse senedi}
{english=stock}
{turkishSemantics=0348790}
{englishSemantics=13438244-n})
```

4.6 Annotation Evaluation

In this current work, all Turkish and English words in the input sentences have been disambiguated by human annotators, who are graduate students in language departments. They are native speakers of Turkish and advanced users of English.

For the evaluation of the annotated dataset, we used an inter-annotator agreement measure. Two different groups of annotators annotated the same sentences. Due to time limitations, we could re-annotate only 500 sentences from both Turkish and English partitions. We got %77.0 and %77.4 of inter-annotator agreement for Turkish and English, respectively.

Table 5: Distribution of sense annotations per synset

(a) Turkish		(b) English	
# of sense annotations	# of synsets	# of sense annotations	# of synsets
(500-1200)	6	(500-665)	2
(300-499)	11	(300-499)	3
(200-299)	15	(200-299)	4
(100-199)	42	(100-199)	22
(50-99)	128	(50-99)	72
(40-49)	53	(40-49)	34
(30-39)	108	(30-39)	79
(20-29)	200	(20-29)	141
(10-19)	521	(10-19)	478
(5-9)	898	(5-9)	921
4	491	4	494
3	529	3	694
2	1524	2	1678
1	2443	1	4037

5 Statistics About the Corpus

5.1 Distribution of Sense Annotations

Except the unambiguous entities, the current status of the Turkish side of the corpus contains 59,847 sense annotations. There are 6,969 distinct sense annotations and the average number of samples per sense is 8.59. The distribution of sense annotations per synset is given in Table 5(a).

For the English partition of the corpus, only entities residing in PWN are annotated, which include nouns, verbs, adjectives and adverbs. The current status of the English partition of the corpus contains 41,986 sense annotations. There are 8,629 distinct sense annotations and the average number of samples per sense is 4.87. The distribution of sense annotations per synset is given in Table 5(b).

5.2 Missing Annotations

When we compare annotations on the English partition with the annotations on the Turkish side, we see that, for some words in English, there is no corresponding semantic annotation in Turkish.

In total, there are 1,323 such words in English, composed of mostly modals (a total of 534: 100 “were”, 209 “was”, 7 “have”, 9 “has”, 6 “had”, 32 “been”, 16 “be”, 155 “are”) and prepositions (a total of 457: 13 “a”, 10 “about”, 2 “around”, 17 “as”, 36 “at”, 12 “back”, 2 “before”, 21 “down”, 5 “even”, 20 “for”, 30 “in”, 12 “into”, 15 “no”, 61 “not”, 22 “of”, 17 “off”, 14 “on”, 53 “out”, 11 “over”, 6 “through”, 4 “to”, 65 “up”, 9 “well”).

5.3 Multiword Expressions

Not only some words on the English partition may have multiword expression counterparts on the Turkish side, but also there are multiword expressions on the English partition whose counterparts are also multiword expressions on the Turkish side. The annotation framework can detect multiword expressions consisting of two and three word expressions (See Section 4.3.2). In total, there are 3,911 two-word (1,215 distinct) and 29 three-word (18 distinct) annotated multiword expressions.

6 Conclusion

In this paper, we reported our experience on manual tagging of English and Turkish senses in an English-Turkish parallel treebank, which had been parsed and enhanced with morphological features before the semantic annotation process. Our study has shown that it is possible to perform a parallel semantic annotation for an English-Turkish corpus and that the pre-processing stage for the parsing and morphological enhancement has been useful as it has accelerated the sense annotation process by providing the annotators with shorter lists of senses of a word in a given sentence.

As a future work, we plan to expand the size of the corpus by following the same manner of procedure, perform word sense disambiguation experiments on it with various classifiers and feature sets and make use of our parallel corpora in various NLP tasks including automatic annotation, statistical machine translation or word sense disambiguation.

Acknowledgment

This work was supported by Tübitak project 116E104.

References

L. Bentivogli, E. Pianta, and M. Ranieri. 2005. Multisemcor: an English Italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, page 90, Trento, Italy, February.

H. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18.

Razieh Ehsani, Ercan Solak, and Olcay Taner Yildiz. 2018. Constructing a wordnet for turkish using manual and automatic annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3):24.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis, Philological Society*, pages 1–32.

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

N. Ide. 2012. Multimasc: An open linguistic infrastructure for language research. In *Fifth Workshop on Building and Using Comparable Corpora*, Istanbul.

D. Jurgens and I. Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *7th International Workshop on Semantic Evaluation*, Atlanta, Georgia.

C. Leacock, G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265, Princeton, NJ.

M. Lupu, D. Trandabat, and M. Husarciuc. 2005. A Romanian semcor aligned to the English and Italian multisemcor. In *1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School*, pages 20–27, EUROLAN, Cluj-Napoca, Romania.

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308, Stroudsburg, PA, USA.

R. Navigli, D. Jurgens, and D. Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (SEM 2013)*, pages 14–15, Atlanta, Georgia.

H. T. Ng and H. B. Lee. 1996. 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*, pages 40–47, Santa Cruz, CA, USA.

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

Z. Orhan, E. Çelik, and N. Demirgüç. 2007. Turkish lexical sample task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 59–63, Prague, Czech Republic.

T. Petrolito and F. Bond. 2014. A survey of wordnet annotated corpora. pages 236–245, 01.

J. I. Saeed. 1997. *Semantics*. Blackwell.

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

B. Snyder and M. Palmer. 2005. The English all-words task. In *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

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

L. Tan and F. Bond. 2012. Building and annotating the linguistically diverse ntu-mc (ntu-multilingual corpus). *International Journal of Asian Language Processing*, 22:161–174.

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

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