Comparison of Turkish Proposition Banks by Frame Matching

Koray Ak
Department of Computer Engineering
Işık University
İstabul, Turkey
koray.ak@isik.edu.tr

Özge Bakay
Department of Linguistics
Boğaziçi University
İstanbul, Turkey
ozge.bakay@boun.edu.tr

Olcay Taner Yıldız
Department of Computer Engineering
Işık University
İstabul, Turkey
olcaytaner@isikun.edu.tr

***Abstract*—By indicating semantic relations between a predicate and its associated participants in a sentence and identifying the role-bearing constituents, SRL provides an extensive dataset to understand natural languages and to enhance several NLP applications such as information retrieval, machine translation, information extraction, and question answering. The availability of large resources and the development of statistical machine learning methods have increased the studies in the field of SRL. One of the widely-used semantic resources applied for multiple languages is PropBank. In this paper, PropBanks applied for Turkish are compared by checking semantic roles in the frame files of matched verb senses. As this integrated lexical resource for Turkish is aimed to be used in a multilingual resource along with English, creation of an inclusive lexical resource for Turkish is of great importance.**

***Index Terms*—Turkish propbank**

I. Introduction

Semantic role labeling (SRL) is a well-defined task that guides us to understand event characteristics by identifying semantic roles of the words in a sentence. Role-bearing constituents are simply detected by answering questions "who" did "what" to "whom," "where," "when," and "how." These questions are addressed to the predicate of the sentence which represents "what" took place in the event. Extracted semantic information with these questions assists NLP applications such as machine translation, information extraction, and question answering. Recent developments of statistical machine learning methods enable NLP applications to learn complex linguistic knowledge, and help the creation of various semantic resources. Recent semantic resources which provide input for developing statistical approaches are FrameNet [1], PropBank [2] [3], [4], and NomBank [5]. These resources provide a stable semantic representation to understand language structure.

PropBank, the bank of propositions, is a corpus annotated with predicate-argument information and it lists the semantic roles or arguments that each verb can cover. All verbs of Penn Treebank Wall Street Journal [WSJ] news corpus have been annotated with semantic roles. Since corpus is taken from WSJ, its content is diverse and reliable and the collection itself has been a valuable language resource for linguistic research. PropBank has been applied for different languages and widely used by research communities.

PropBank uses the same naming convention for Agent (Arg0) and Theme/Patient (Arg1) across verbs. The other set of arguments can vary from Arg2 to Arg5 which can represent different meanings for different verbs. They can refer to the instrument, starting point, end point, beneficiary, or attribute. Moreover, PropBank uses ArgMs as modifier labels where the role is not specific to the verb group and generalizes them over the corpora such as location, temporal, purpose, or cause arguments etc. Argument roles are named and stored in framesets.

Prior to annotation, verbs of the corpora are analysed and frame files are created as stated in the Framing guidelines of PropBank [6]. Each verb has a frame file which contains arguments applicable to that verb. Frame files provide all possible semantic roles and also, all possible syntactic constructions are represented with examples. Frame files may include more than one roleset for different senses of polysemous verbs. In the roleset of a verb sense, argument labels Arg0 to Arg5 are described with the meaning of the verb. For instance, Figure 1 presents the roles of the predicate "attack" from PropBank, Arg0 is "attacker", Arg1 is "entity attacked", and Arg2 is "attribute".

Roleset id: attack.01 , *to make an attack, criticize strongly,*

attack.01: Member of Vncls judgement-33.

Roles:

Arg0-PAG: *attacker* (vnrole: 33-agent)
Arg1-PPT: *entity attacked* (vnrole: 33-theme)
Arg2-PRD: *attribute*

Fig. 1. Roleset attack.01 from English PropBank for the verb "attack" which includes Arg0, Arg1 and Arg2 roles.

This paper is organized as follows: Since we compare senses in two Turkish PropBanks, we provide information about these resources in Section II, and related work about combinations of English resources in Section III. We give the comparison details in Section IV, and statistics about this comparison in

Section V. Lastly, we conclude in Section VI.

II. Turkish Proposition Banks

There have been different attempts to construct Turkish PropBank in the literature. Şahin [7], [8], Şahin and Adalı [9] report the semantic role annotation of arguments in the Turkish dependency treebank. They construct PropBank by using ITU-METU-Sabancı Treebank (IMST) [10] and later align it with IMST Universal Dependencies (UD). IMST is a syntactically-annotated corpus with sentences from Metu Turkish Corpus, which includes modern Turkish texts from 10 distinct genres. Şahin and Adalı frame 772 verbs and 1285 verb senses. They suggest using morpho-semantic features to calculate frame files for the verbs derived from verb roots by applying valency patterns. They use the Turkish root verbs provided by the Turkish Language Association (TDK) and the Turkish National Corpus (TNC).

In the frame creation process, they use Cornerstone [11], an open source Propbank frameset editor. They adjust Cornerstone for Turkish by adding two dropdown menus to supply case marking information and possible suffixes for Turkish verbs. Verb sense disambiguation of the frames and argument annotation are completed by using crowd sourcing. 20060 semantic roles are annotated in 5635 sentences. Validity and quality control process of crowd sourcing techniques are discussed and results are evaluated in their paper.

Recently, Ak et al. [12] have constructed another Turkish Proposition Bank using translated sentences of English PropBank. 9560 sentences containing a maximum of 15 tokens from the translated Penn Treebank II [13], [14] are used to generate this proposition bank. This corpus consists of 8660 sentences from the training set of the Penn Treebank [15], 360 from its development set, and 540 from its test set. According to the paper, 9560 of 17000 translated sentences are annotated with semantic roles. Along with the annotated corpus, framesets are created for 1914 verb senses. The Penn Treebank structure provides advantages for building a fully-tagged data set including syntactic labels, morphological labels, and parallel sentences in English and Turkish.

Ak et al. use an in-house-developed freely-available toolkit, NLP Toolkit, both for frame creation and annotation processes. This tool also includes other NLP methods, for preprocessing words in the dataset such as morphological and semantic analyses for verb sense selection. The tool is used by multiple annotators for frame creation and annotation of the words by hand. Frames are stored in a single file where each verb sense has a frameset. Framesets are distinguished by a unique synset id, e.g. "TUR10-0006410". Synset ids are taken from a Turkish WordNet based on the Contemporary Dictionary of Turkish provided by TDK.

III. Related Work

WordNet is a graph data structure where the nodes are word senses with their associated word forms and edges are semantic relations between the sense pairs. The first WordNet project was Princeton WordNet (PWN), which was initiated by George Miller in 1995, [16]. Over time, PWN has evolved to become a comprehensive relational representation of the word senses of English [17].

SemLink [18] is an integrated resource combining information provided by various lexical resources (VerbNet [19], FrameNet [20], PropBank [4] and WordNet [17]). Through mappings among these resources, the project aims to develop an NLP resource with an extended overall coverage.

By focusing on predicates, López de Lacalle et al. ([21]–[24]) develop a common semantic infrastructure, called the Predicate Matrix (PM), with the aim of having interoperability among the same resources used in SemLink. Some of the methods they use to achieve this integration automatically are graph-based word sense disambiguation algorithms and various corpus alignment methods. Criticising the limitations of the manual methods used in developing SemLink, they conduct the integration process automatically and suggest the use of automatic integration for a more inclusive semantic processing.

IV. PropBank Comparison

In order to compare framesets, a mapping between proposition banks is required. In the subsection IV-A, frameset mapping process is described, and then in subsection IV-B comparison procedure is presented in detail.

A. Frame Matching

Turkish proposition banks discussed in Section II are built on different datasets, even frame file standards differ from each other. However, both of them use the verbs from the dictionary of Turkish Language Association. As we state in the previous section, root verbs of the first study is taken from TDK as in the second study and the synset ids are taken from WordNet based on the TDK dictionary. So we gathered frame files and WordNet to match verb senses. Şahin serves their frame files from the Turkish PropBank github website. We downloaded frame files and parsed each frame file to collect verb senses. In Figure 2, sample frame file for "bekle" is presented. This predicate has three senses stored in "roleset" tags. In each tag, the "name" attribute contains the meaning of the verb sense.

Likewise, we downloaded the WordNet file from the website [25] and the frame files from the website [12] for the proposition bank of Ak et al. Since rolesets use synset ids from WordNet, it is sufficient to use WordNet to match verb senses. In Figure 3, a part of a frame file is presented for "bekle" where WordNet ids are stored in the "id" attribute.

WordNet entry for "TUR10-0089560" is shown in Figure 4. The fourth sense of "Beklemek" in the WordNet has the same meaning as bekle.02 in the Turkish PropBank of Şahin.

We automated this process to match verb senses across the proposition banks. The frame matching method simply creates a list of verb sense ids along with the corresponding meanings for both proposition banks. If there is any verb sense with the exactly same meaning, our method pairs the ids at the top for this verb. For unmatched verb senses, meanings are listed for both sides. After we generated the verb sense lists, linguists in

TABLE I
BEFORE THE MATCHING

Verb	Example sentence	PropBank Senses	WordNet Senses
bağışlamak	Çocuk elindeki çiçek demetini kumandanın ayağı altına atarak, babamı bağışlayınız, diyordu.	Hoş görmek, affetmek	Herhangi bir kötü davranış için ceza vermekten vazgeçmek; affetmek
	Ödünç aldığı parayı bile kendinden daha ihtiyaçlısına bağışlayan ancak bir masal adamıdır.	Bir mal veya hakkı karşılık beklemeden birine vermek	Görevden, işten çekmek
			Deyimlerde "Tanrı esirgesin, ayırmasın" anlamlarında kullanılan bir söz
			Bir mal veya hakkı karşılık beklemeden birine vermek, teberru etmek

TABLE II
AFTER THE MATCHING

Verb	Example sentence	PropBank Senses	WordNet Senses
bağışlamak	Çocuk elindeki çiçek demetini kumandanın ayağı altına atarak, babamı bağışlayınız, diyordu.	Hoş görmek, affetmek	Herhangi bir kötü davranış için ceza vermekten vazgeçmek; affetmek
	Ödünç aldığı parayı bile kendinden daha ihtiyaçlısına bağışlayan ancak bir masal adamıdır.	Bir mal veya hakkı karşılık beklemeden birine vermek	Bir mal veya hakkı karşılık beklemeden birine vermek, teberru etmek

```
▼<frameset>
 ▼<predicate lemma="bekle">
  ▼<roleset id="bekle.01" name="Birini ya da birşeyi beklemek" vncls="">
   ▼<roles>
    ▼<role descr="bekleyen kişi" f="" n="0" suffix="NOM">
       <vnrole vncls="" vntheta="agent"/>
     </role>
    ▼<role descr="beklenilen kişi/şey" f="" n="1" suffix="ACC">
       <vnrole vncls="" vntheta="theme"/>
     </role>
     <note/>
   </roles>
   ▼<example name="tdk.01" src="" type="">
     <text>Ben de seni bekliyordum zaten.</text>
     <arg f="" n="0" suffix="">Ben de</arg>
     <arg f="" n="1" suffix="">seni</arg>
     <arg f="adv" n="m" suffix="">zaten</arg>
     <note/>
   </example>
   <note/>
  </roleset>
 ▶<roleset id="bekle.02" name="Ummak" vncls="">...</roleset>
 ▶<roleset id="bekle.03" name="Korumak">...</roleset>
  <note/>
 </predicate>
 <note/>
</frameset>
```

Fig. 2. Frame file for “bekle” taken from Turkish PropBank github web page.

```
<FRAMESET id="TUR10-0089650">
    <ARG name="ARG1">Beklenen</ARG>
    <ARG name="ARG0">Bekleyen</ARG>
    <ARG name="ARGMTMP">Bekleme zamanı</ARG>
</FRAMESET>
<FRAMESET id="TUR10-0089780">
    <ARG name="ARG1">Beklenen</ARG>
    <ARG name="ARGMDIS">Söylem işaretçisi</ARG>
    <ARG name="ARG2">Beklenilen şey/kişi</ARG>
    <ARG name="ARGMTMP">Beklenme zamanı</ARG>
</FRAMESET>
```

Fig. 3. Frame file of Ak et al. [12].

the team matched the verb senses manually. For instance, for the verb “bağışlamak”, the linguists were provided with a list

```
<SYNSET>
    <ID>TUR10-0089650</ID>
    <SYNONYM>
        <LITERAL>beklemek<SENSE>4</SENSE></LITERAL>
        <LITERAL>demek<SENSE>8</SENSE></LITERAL>
        <LITERAL>umut etmek<SENSE>1</SENSE></LITERAL>
    </SYNONYM>
    <POS>v</POS>
    <ILR>ENG31-00721987-v<TYPE>SYNONYM</TYPE></ILR>
    <DEF>Ummak</DEF>
    <EXAMPLE>Nikâhtan bu kadar keramet bekleme!</EXAMPLE>
</SYNSET>
```

Fig. 4. WordNet synset for Frameset id “TUR10-0089560”.

as in Table I including two senses from Şahin’s PropBank and four from WordNet. Additionally, two exemplary sentences were given to make the distinction among the senses clearer. After analyzing the senses in PropBanks, the corresponding senses were matched as it can be seen in Table II.

B. Frame Comparison

After matching proposition bank of Şahin and WordNet, we composed a mapping file which consists of a WordNet id and an equivalent roleset id as in Table III. As we previously stated in Subsection IV-A and Figure 2, the “id” attribute of “roleset” tag contains the necessary information to find the corresponding WordNet Id to match the rolesets of the two Turkish proposition banks. So, for each frame file of Şahin, we listed rolesets and found the corresponding WordNet ids from the mapping. Then, we searched for the frameset of the WordNet id in Ak et al’s frame file. Once we found a frameset, we compared the arguments one by one.

V. RESULTS

After examining 1285 senses of 772 verbs in total, it has been observed that the majority of the senses (1111 senses of

TABLE III
MAPPING BETWEEN WORDNET AND PROPBANK (ŞAHIN)

RolesetId	WordNetId
aban.01	TUR10-0000360
abartıl.01	TUR10-0000500
abart.01	TUR10-0129480
acık.01	TUR10-0002820
acı.01	TUR10-0002890

711 verbs) in the PropBank(Şahin) and WordNet do match. However, although both datasets were created by using the items in TDK (the Turkish Language Association), there are differences between them. The reason behind those differences can be that they were created by using different versions of TDK.

The first difference between the datasets is that in addition to those 772 verbs that were compared, there are 1322 verb senses which do not exist in Şahin's PropBank and therefore, are not included in the comparison.

The second difference is that 170 senses in Şahin's PropBank dataset do not have matching senses in WordNet.

Apart from verbs, there were also 4 frame files for the suffixes "-da", "-la", "-lan", and "-laş", which add new roles to the verbs that they produce. We also omitted these frames since they have multiple correspondents in WordNet.

The last difference is that 4 senses of 4 verbs in WordNet were merged in Şahin's and thus, one-to-one match between those senses was not possible. For example, for the verb "bulmak" (to find), one of the senses provided in WordNet was "keşfetmek, icat etmek" (to discover, to invent). However, this sense was split into two in Şahin's: "Varlığı bilinmeyen bir şeyi ortaya çıkarmak|Var olduğu bilinmeyen bir şeyi bulmak" (to discover) and "İlk kez yeni bir şey yaratmak" (to invent). The other three verbs whose senses were merged in Şahin's are "bulunmak", "hırpalamak", and "utanmak".

In frame and roleset comparison, we have tried to compare the roles of 1111 verb senses that are mapped to WordNet. However, the rolesets for 519 verb senses do not exist in the frame file of Ak et al., which is equal to half of the verb senses from the intersection of WordNet and frames of Şahin. So we have only compared 592 common verb senses in this step.

When we process the rest of the rolesets, we have observed that the rolesets of Şahin's frames have 2713 roles for 1285 verb senses where 372 roles are added for the 170 verb senses which do not exist in WordNet. Also, 1050 roles are added for 519 verb senses not exist in Ak et al's study. There are also 430 roles that exist in Şahin but not exist in Ak et al. On the other hand, Ak et al. offers 1568 roles for 592 verb senses in common. Within the common set, 861 roles are the same with the roles of Şahin. Remaining 707 roles do not exist in Şahin's frames. When we investigate further, we have found out that access modifiers such as temporal, locative, purpose, and cause are added along with the ARG0-ARG5 arguments to the rolesets in Ak et al. In Şahin's frames, these modifier roles are not included, they only add argument from ARG0 to ARG5. Table IV lists number of roles that are same in both proposition bank. In Table V number of roles that exist in Şahin but not exist in Ak et al displayed. And in VI number of roles that exist in Ak et al but not exist in Şahin presented. As you can see, ARG0 and ARG1 arguments are the most common roles that are same in both propbanks. The main difference for propbanks is access modifiers treated differently in both propbank.

TABLE IV
NUMBER OF SAME ROLES IN THE ROLESET

Argument	Count
ARG0	404
ARG1	441
ARG2	16

TABLE V
NUMBER OF ROLES IN ŞAHIN DIFFERS FROM AK ET AL.

Argument	Count
ARG0	86
ARG1	90
ARG2	158
ARG3	48
ARG4	44
ARG5	1
ARGm	3

TABLE VI
NUMBER OF ROLES IN AK ET AL DIFFERS FROM ŞAHIN.

Argument	Count
ARG0	82
ARG1	39
ARG2	24
ARG3	8
ARGMADV	36
ARGMCAU	14
ARGMDIR	6
ARGMDIS	90
ARGMEXT	54
ARGMLOC	80
ARGMMNR	88
ARGMPNC	41
ARGMTMP	145

VI. CONCLUSION

In this paper, we present the comparison between and the matching process of the newly-created Turkish Proposition Banks. As Turkish is not very rich in terms of linguistic resources, relating these two propbank studies may give the opportunity to improve both datasets. Frame files that do not exist in both sides can be imported or role differences for matched verb senses can be reevaluated.

There are 2397 rolesets and 6090 roles for these rolesets in the study of Ak et al. and Şahin's frame files consist of 1285 rolesets and 2713 roles defined in frame files. Also, mapping these proposition banks with WordNet is important. Turkish WordNet used in the mapping is aligned with the English

counterpart, which may enable the transfer of linguistic information from other proposition banks and give possibility to extend datasets.

ACKNOWLEDGMENT

This work was supported by Tübitak project 116E104.

REFERENCES

[1] C. J. Fillmore, J. Ruppenhofer, and C. F. Baker, *FrameNet and Representing the Link between Semantic and Syntactic Relations*, ser. Language and Linguistics Monographs Series B. Taipei: Institute of Linguistics, Academia Sinica, 2004, pp. 19–62.

[2] P. Kingsbury and M. Palmer, "From treebank to propbank." in *LREC*. European Language Resources Association, 2002.

[3] ——, "Propbank: The next level of treebank," in *Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden, 2003.

[4] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Comput. Linguist.*, vol. 31, no. 1, pp. 71–106, Mar. 2005.

[5] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman, "The nombank project: An interim report," in *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 24–31.

[6] O. Babko-Malaya, *Guidelines for Propbank Framers*, 2005.

[7] G. G. Şahin, "Framing of verbs for turkish propbank," in *TurCLing 2016 in conj. with 17th International Conference on Intelligent Text Processing and Computational Linguistics*, 2016.

[8] ——, "Verb sense annotation for turkish propbank via crowdsourcing," in *17th International Conference on Intelligent Text Processing and Computational Linguistics*, 2016.

[9] G. G. Şahin and E. Adalı, "Annotation of semantic roles for the Turkish proposition bank," *Language Resources and Evaluation*, May 2017.

[10] U. Sulubacak, T. Pamay, and G. Eryiğit, "Imst: A revisited Turkish dependency treebank," in *The First International Conference on Turkic Computational Linguistics*, 2016, pp. 1–6.

[23] M. L. de Lacalle, E. Laparra, I. Aldabe, and G. Rigau, "A multilingual predicate matrix," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Paris, France, may 2016.

[11] J. D. Choi, C. Bonial, and M. Palmer, "Propbank frameset annotation guidelines using a dedicated editor, cornerstone." in *LREC*, 2010.

[12] K. Ak, O. T. Yıldız, V. Esgel, and C. Toprak, "Construction of a Turkish proposition bank," *Turkish Journal of Electrical Engineering and Computer Science*, vol. 26, pp. 570 – 581, 2018.

[13] O. T. Yıldız, E. Solak, O. Görgün, and R. Ehsani, "Constructing a Turkish-English parallel treebank," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, 2014, pp. 112–117.

[14] O. T. Yıldız, E. Solak, Ş. Çandır, R. Ehsani, and O. Görgün, "Constructing a Turkish constituency parse treebank," in *Information Sciences and Systems 2015*. Springer, 2015, pp. 339–347.

[15] 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.

[16] G. A. Miller, "Wordnet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[17] C. Fellbaum, "ed. wordnet: an electronic lexical database," *MIT Press, Cambridge MA*, vol. 1, p. 998, 1998.

[18] M. Palmer, "Semlink: Linking propbank, verbnet and framenet," in *Proceedings of the Generative Lexicon Conference*, Pisa, Italy, 2009.

[19] K. Kipper-Schuler, "VerbNet: a broad-coverage, comprehensive verb lexicon," Ph.D. dissertation, Computer and Information Science Department, University of Pennsylvania, Philadelphia, PA, 2005.

[20] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley FrameNet project," in *COLING-ACL '98: Proceedings of the Conference*, Montreal, Canada, 1998, pp. 86–90.

[21] M. L. de Lacalle, E. Laparra, and G. Rigau, "First steps towards a predicate matrix," in *Proceedings of the 7th Global WordNet Conference*, Tartu, Estonia, Jan 25-29 2014.

[22] ——, "Predicate matrix: extending semlink through wordnet mappings," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2014.

[24] ——, "Predicate matrix: automatically extending the semantic interoperability between predicate resources," *Language Resources and Evaluation*, vol. 50, no. 2, pp. 263–289, Jun 2016.

[25] R. Ehsani, E. Solak, and O. T. Yıldız, "Constructing a wordnet for Turkish using manual and automatic annotation," *ACM Transactions no Asian Low-Resource Language Information Processing*, vol. 17, no. 3, 2018.