TRopBank: Turkish PropBank V2.0

Neslihan Kara1, Deniz Baran Aslan1, Büşra Marşan1, Özge Bakay2, Koray Ak3, Olcay Taner Yıldız2

Starlang Yazılım Danışmanlık1, Boğaziçi University2, Akbank3, Işık University4

neslihan@starlangyazilim.com, deniz@starlangyazilim.com, busra@starlangyazilim.com, ozge.bakay@boun.edu.tr, koray.ak@akbank.com, olcaytaner@isikun.edu.tr

Abstract

In this paper, we present and explain TRopBank “Turkish PropBank v2.0”. PropBank is a hand-annotated corpus of propositions which is used to obtain the predicate-argument information of a language. Predicate-argument information of a language can help understand semantic roles of arguments. “Turkish PropBank v2.0”, unlike PropBank v1.0, has a much more extensive list of Turkish verbs, with 17,673 verbs in total.

Keywords: Propbank, Turkish PropBank, Argument Structure Annotation

1. Introduction

Verbs constitute a major category in human languages, expressing the critical information concerning a state or an event. However, having the grasp of the mere definition of this category is not sufficient for comprehending the meaning or function of a given verb within a sentence. In order to do so, another essential component of verbs must be introduced: the argument structure. By placing a verb within the proper grammatical context and associating it with its arguments, any verbal structure can be analyzed accurately. With PropBank, our aim is to provide this indispensable contextual information through annotating the argument structure of each verb. Thus it is evident that PropBank’s function is indispensable for processing and properly interpreting Turkish. In addition, PropBank enhances numerous NLP applications (e.g. machine translation, information extraction, question answering and information retrieval) by adding a semantic layer to the syntax, which takes the whole structure one step closer to human language.

Being the complements of a verb, arguments express grammatical information that is classified in accordance with their syntactic and semantic roles. In a sentence like "Jack gave Jenny a present", the verb “give” has a structure which corresponds to a list of arguments. "Jack" is the subject (agent), "a present" is the direct object (theme) and "Jenny” is the indirect object (recipient). Note that each verb has a different argument structure and requires a different number of arguments in various semantic roles. With TRopBank annotations, certain liberties have been taken in order to produce a more comprehensive corpus, where non-obligatory information has also been included as arguments. In theoretical syntax, non-obligatory bits of information are classified as adjuncts, contrasting with obligatory arguments. Nonetheless, with TRopBank, the scope of the term “argument” has been kept as wide as possible in order to provide an accurate representation of thematic roles.

In this paper, we present our approach in expanding Turkish PropBank. The structure of this paper is as follows: In Section 2., we review the literature in order to provide information about PropBanks created for other languages. Section 3. presents details regarding the structure of verbs in Turkish. Section 4. gives information on the annotation process we followed, the problems encountered during this process and their respective solutions. Section 5. offers some statistics regarding the annotated verbs and a compendia commentary. Lastly, Section 6. concludes the paper with final remarks.

2. Literature Review

The link between syntactic realization and semantic roles was mentioned in Levin’s comprehensive study (Levin, 1993). Syntactic frames, which are diagrammatic representations of events, were stated as a direct reflection of the underlying semantics and associated with Levin classes, which define the allowable arguments for each class member. VerbNet (Kipper et al., 2000) extends these classes that were defined by Levin. In VerbNet, abstract representation of syntactic frames for each class was added to Levin classes. These representations include explicit correspondences between syntactic positions and semantic roles. For example for “break” Agent REL Patient, or Patient REL into pieces added. FrameNet (Fillmore et al., 2004) is another semantic resource based on Frame Semantics theory. FrameNet proposes semantic frames to understand the meaning of most words and these semantic frames include a description of a type of event, relation, entity and participants. For example, the concept of cooking typically involves an agent doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food (Container) and a heat source. Another semantic resource is PropBank (Kingsbury and Palmer, 2002) (Kingsbury and Palmer, 2003), (Palmer et al., 2005) (Bonial et al., 2014) which includes predicate-argument structure by stating the roles that each predicate can take along with the annotated corpora. Prior to PropBank annotation, frame files were constructed to include possible arguments for verbs or nouns. These frame files help users label various arguments and adjuncts with roles.

Studies for the construction of the English PropBank date back to 2002. In the first version of the English PropBank, annotation effort focused on the event relations that are ex-
pressed only by verbs. Prior to annotation, verbs of the corpora were analysed and frame files were created. Each verb has a frame file which contains arguments applicable to that verb. Frame files provide all possible semantic roles as well as all possible syntactic constructions, which are represented with examples. In the roleset of a verb sense, argument labels Arg0 to Arg5 are described with the meaning of the verb. Figure 1 presents the roles of predicate. The roles of predicate “attack” are as follows; 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 Tulsa judgement-33.

Roles:

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

Figure 1: Roleset attack.01 from English PropBank for the verb “attack” which includes Arg0, Arg1 and Arg2 roles. PAG = agent, PPT = theme, PRD = predication

In most of the rolesets, two to four numbered roles exist. However, in some verb groups, such as verbs of motion, there can be six numbered roles in the roleset. In the frame construction phase, numbered arguments are selected among the arguments and adjuncts in the sentence. Most of the linguists consider any argument higher than Arg2 or Arg3 to be an adjunct. In PropBank, if any argument or adjunct occurs frequently enough with their respective verbs, or classes of verbs, they are assigned a numbered argument to ensure consistent annotation. Arg2 to Arg5 labels in the frame files may indicate different roles for the different senses of the verb. On the other hand, similar roles are assigned for Arg2 to Arg5 for the verbs in the same Levin class. For example buy, purchase and sell are in the same Levin class. The rolesets for buy and purchase are the same and they are similar to sell rolesets since Arg0 role of the first group is equivalent to Arg2 role of the sell roleset. Rolesets of these verbs are represented below in Table 1.

Table 1: Rolesets buy, purchase and sell from English PropBank consist of the same roles.

| PURCHASE       | BUY       |
|----------------|-----------|
| ARG0: buyer    | ARG0: buyer |
| ARG1: thing bought | ARG1: thing bought |
| ARG2: seller   | ARG2: seller |
| ARG3: price paid | ARG3: price paid |
| ARG4: beneactive | ARG4: beneactive |

SELL

| ARG0: seller |
| ARG1: thing sold |
| ARG2: buyer |
| ARG3: price paid |
| ARG4: beneactive |

(1) John broke the window (causative),(transitive)

(2) The window broke (inchoative),(intransitive)

Figure 2: Break in both inchoative and causative constructions.

Semantic role annotation begins with a rule-based automatic tagger, and afterwards the output is hand-corrected. Annotation process is straight-forward; whenever a sentence is annotated, annotators select the suitable frameset with respect to the predicate and then tag the sentence with the arguments that are provided in the frameset file. Syntactic alternations which preserve verb meanings, such as the causative and inchoative alternation or object deletion, are considered to be one frameset only. Annotators start with Arg0 to the annotation since any argument satisfying two or more roles should be tagged with the highest ranked argument where the priority goes from Arg0 to Arg5. PropBank also offers solutions to annotation disagreements by adopting double-blind annotations to increase the quality of the annotation. Whenever a disagreement occurs between the annotators, an adjudicator decides the correct annotation and new roles may be added to the roleset. Semantic information annotated in the first version of the PropBank is based solely on verbal predicates. Generally verbs provide the majority of the event semantics of the sentence. However, to extract complete semantic relations of the event, new predicate types such as nouns, adjectives
and complex predicate structures like light verbs should be taken into account. These new predicate types are included in the latest version of the PropBank, which offers guidelines specific to each structure that bears semantic information about the event.

Via different syntactic parts of speech, identical events can be expressed differently. Figure 3 gives examples for the same events with different syntactic parts of speech. In the first example, fear of mice is represented with verb, noun and adjective forms and gives the same semantic information about the event. The second example with "offer" also concludes the same semantic meaning across verb, noun and multi-word constructions of the word. Semantic information is already covered for noun, adjective and complex predicates in FrameNet but PropBank expands its coverage to new predicate types. For the nominal frame files, PropBank relied on the NomBank in the initial creation of frames. Among all the noun types in NomBank, only the eventive nouns were processed in PropBank. Also, WordNet and FrameNet are visited to expand PropBank’s nominal and adjective frame files coverage, and to assess derivational relationships between new predicate type role sets.

She fears mice. OR She is afraid of mice. OR She is afraid of mice.

Figure 3: Different syntactic constructions of the same event.

In the previous version of PropBank, adjectives followed by copular verbs, as in the first example in Figure 3, are annotated with respect to the semantics of the copular verb. Annotation of the example sentence with respect to the previous version of PropBank is shown in Figure 4. As can be seen, annotation with respect to the verbal predicate in this sentence does not reveal the complete semantic meaning. A fearing event is not understood from the annotation. The reason of incomplete semantic representation is the adjectives in this kind of sentences having more semantic information than the verbal predicates. To overcome this, annotation has expanded to include predicate adjectives in the new version.

Relation: is
Arg1-Topic: She
Arg2-Comment: afraid of mice

Figure 4: Annotation of the sentence with respect to the copular verb in the previous version of PropBank.

The annotation of the same sentence with respect to predicate adjectives gives the result in Figure 5. Although the bulk semantic information is based on adjectives in this kind of sentences, the copular verb "to be" does play a role in the sentence and annotation of the copular verb is also required for complete semantic representation. The subject of the adjectival predicate is syntactically an argument of the copular verb rather than an argument of adjectival one afraid. To gather all the event participants in the sentence, PropBank annotates copular verbs and their adjectival domain, which contains the experiencer argument. Then it re-annotates the sentence with respect to the adjectival predicate and its syntactic domain.

She is afraid of mice.

Rel: is afraid
Arg0: She
Arg1: of mice

Figure 5: Annotation of the sentence with respect to predicate adjective.

Furthermore, PropBank recently added eventive and stative nouns which occur inside or outside the light verb constructions to the focus of annotation. In the initial phase, more than 2,800 noun role sets are added to the frame files. Most of these role sets are taken from NomBank frames, and the coverage is expanded using WordNet definitions which state the noun types as noun.event, noun.act, noun.state for the eventive and stative nouns. Similar to adjectival predicates, verbs in complex predicates, such as the ones in light verb constructions, are annotated with their syntactic domains; then annotation for the noun part is processed i.e. the light verb construction make an offer is annotated for both make and offer.

ARG0: entity offering
ARG1: commodity, thing offered
ARG2: price
ARG3: benefactive or entity offered to

[Yesterday]ARGM-TMP, [John]ARG0 [made]REL an [offer]REL to [buy the house]ARG1 [for $350,000]ARG2.

Figure 6: Annotation of the sentence with respect to noun in the LVC.

In the first version of the PropBank, noun light verb construction (LVC) is ignored and the situation is handled by using either one of the role sets of the dominant sense of the verb or a designated role set for the LVC. As a result, semantic information that is presented by the noun is omitted. In the current version, annotators identify the light verbs and main nominal predicate in the first pass, then annotation is done with respect to complete arguments of the complex predicate by looking into the role set of nominal predicate. In the example in Figure 6, annotation is completed using the role set of offer and roles for both make and offer are extracted.
2.1. PropBank Studies in Different Languages

Apart from English, PropBank studies have been conducted for several languages. In Figure 7, publications for different languages are presented in a timeline. Unlike the rest, German & Japanese are not annotated in PropBank style. For German, Frame Based Lexicon corpus is annotated in the framework of Frame Semantics. Japanese Relevance Tagged Corpus is annotated for relevance tags such as predicate-argument relations, relations between nouns and coreferences. PropBank style arguments are not used but since predicate-argument relations are tagged, the corpus can be regarded as a proposition bank. Also, argument annotations can be converted to PropBank style with ease.

- **Arabic**: Palmer et al. (2008) have created the Pilot Arabic PropBank, consisting of 200,000 words and 24 label types. They employ frame sets for the annotators’ sake including predicate and its possible arguments since Arabic has a different system for writing and speaking. Also they use *lemmas* for the root of the verbs since derivation happens around lemmas. Later Zaghouani et al. (2010) have revised the Pilot Arabic PropBank. They have reviewed and added new Frame Files: at the end, all lemmas have their own Frame Files. They have also added gerunds.

- **Basque**: Agirre et al. (2006) present a methodology for adding a semantic layer to the Reference Corpus for the Processing of Basque, a 300,000-word sample collection, applying the PropBank model. Aldezabal et al. (2010a) and Aldezabal et al. (2010b) present their work in adding semantic relation labels to the Basque Dependency Treebank, tagging about 12,000 words of the corpus. They also point out that the bulk of the tagging can be done automatically, leaving only a small portion to be tagged manually.

- **Chinese**: Xue (2006), X. and Palmer (2009) present Chinese PropBank, a semantic lexicon consisting of 11,765 predicates, which is built upon the Chinese Treebank. The annotations include not only arguments but adjuncts as well. The predicates are separated according to their distinct senses and each is assigned a frameset to be filled. Palmer et al. (2008) expand upon previous work and build a Chinese parallel of PropBank II, which adds further semantic information to the annotations.

- **Dutch**: Based on the Dutch corpus SoNaR, in their study, De Clercq et al. (2012) analyze approximately 1 million items in terms of named entities, co-reference relations, semantic roles and spatio-temporal relations. They annotate one half of the data manually, the other half automatically. They conclude that the automatic labeller performs better on verbs with less arguments and for manually annotated data, as it is often hard for annotators to decide on a single meaning for a Dutch verb given an English one.

- **English**: Kingsbury and Palmer (2002), Kingsbury and Palmer (2003) annotate English verbs through a two-tiered action plan. In the first tier, ARG0 and ARG1 labels (abstract labels) in accordance with their verb-sense specific meanings and ARG1 labels are employed for unaccusative verbs whereas ARG0 is preferred for all other verbs. After the first tier is complete, the labels of the second tier are assigned in accordance with their prominence.

- **Finnish**: Haverinen et al. (2015) present the Finnish PropBank, which is built upon the Turku Dependency Treebank, made up of over 15,000 sentences. They use a modified version of the Stanford Dependency scheme. Their workflow breaks down into two stages, where they first create the framesets for the annotation and then use them to annotate each occurrence. They utilize an efficient strategy where the initially-created framesets are applied to categories of verbs that they fit, in batches, with necessary modifications. They also chose to include derived causatives, which have been excluded from Turkish PropBank.

- **French**: van der Plas et al. (2010) annotate 1040 entries from the Europarl corpus within the framework of PropBank. They employ the same 2-tier approach as the English PropBank. First, they mark agents with A0 (arg0) and patients with A1. Second, they assign the appropriate labels (A0, A1, A2, A3, A4 and A5) to the remaining arguments. First, 4 annotators annotate 130 entries manually for training and calibration purposes. Then the remaining 900 entries are annotated by a single annotator. Except for idioms and collocations (130 tokens), their annotations are parallel to those of the English PropBank.

- **German**: Erk et al. (2003) semi-automatically annotate 1320 entries from the TIGER corpus. They use FrameNet and mainly two semantic frames: request frame and commercial transaction frame. They use 7 and 8 fold annotations for the request frame, and 2, 3 and 5 fold annotations for the commercial transaction frame. Burchardt et al. (2006) manually annotate 8,700 lexical units from the TIGER corpus employing semantic frames of operate vehicle, statement, ride vehicle and support. Multi-word idioms are treated as single units and annotated in relation to their meanings.

- **Hindi**: In order to complete the analysis faster, Vaidya et al. (2011) first analyze the similarities between dependency and predicate-argument structures, then match the syntactic dependents with semantic arguments with a rule-based system. They also use the label PRO for empty elements.

- **Japanese**: Kawahara et al. (2002) present the Japanese Relevance-tagged Corpus, so far including thirteen hundred tagged sentences. The sentences are drawn from the Kyoto University Corpus, which consists of 40,000 syntactically tagged sentences. The sentences are tagged in regards to predicate-argument relations and the relations between nouns. The relations are decided depending on surface case, unlike Turkish PropBank where the semantics of the verbs are prioritized.
3. Remarks on the Structure of Turkish Verbs

3.1. Linguistic Evaluation of Turkish Verbs

As an agglutinating language, Turkish has a rich inventory of morphological forms that can attach to word roots in order to modify their category, meaning or grammatical function. In this paper, one category of these morphological forms is of particular interest: voice suffixes. In Turkish, the base form of the verb is the active voice. In order to derive other voices, their respective suffixes are attached to the verb.

The verb acquires passive voice through the attachment of a passivizing suffix. As a result of this operation, the subject is removed and the object is promoted to the subject position. This passivization operation is utilized very frequently in Turkish. Another example is the causative voice, in which the verb acquires a causer, which becomes the subject of the sentence. The former subject is then demoted to the object position. The passive and causative voice suffixes are important in that they can attach to most active verbs (so they are highly productive and rule-governed). Other voices that are overtly marked in Turkish are the reciprocal and reflexive. However, these two voices are very limited in use compared to the rest, as they are applicable to few verbs. In Figure 8, a list of examples with corresponding voice changes are provided.

Note that voice suffixes in Turkish can be stacked. In other words, a passive suffix and a causative suffix can be found on a single verb, or a verb can take two passive suffixes and become a double passive construction. These operations are relevant to TRopBank as they modify a verb’s syntactic structure, reducing or increasing the amount of its arguments. In our annotations, we have decided to exclude all passive suffixes and adverbial arguments. There are 20 of them in total but they do not use ARG5 since noun incorporation is highly common in Persian.

As illustrated in this short review, there are Proposition Banks available in many languages, built upon different corpora using a variety of tools and techniques. While all of these meet the criteria to qualify as PropBanks, they have been built in accordance with different principles, and they include detailed information to differing extents. The available resources display inconsistency in terms of what is to be included and what is to be left out, and the structural differences of the respective languages are no doubt a significant factor.
(3) Ahmet kitabı oku-du. *Ahmet read the book.* (active)

(4) Kitap oku-n-du. *The book was read.* (passive)

(5) Ayşe Ahmet’e kitabı oku-t-tu. *Ayşe made Ahmet read the book.* (causative)

(6) Ayşe ile Ahmet öp-iüş-tü. *Ayşe and Ahmet kissed.* (reciprocal)

(7) Ali yıka-n-di. *Ali washed himself.* (reflexive)

Figure 8: Examples for Voice in Turkish

Arbitrary. This means that they are not rule-governed; therefore, they must be included as independent entries in order to be properly analyzed. Voice suffixes are not the only point of concern for our analysis. Turkish also has a group of auxiliary or light verbs. They are essentially verbs that attach to other verb roots in order to modify their meanings. One example is -(y)Abil, historically derived from the verb bil- “to know”, which adds the meaning of ability, permission, or possibility. This corresponds to the English modal verb “can”. Another common example is -(y)Iver, which adds the meaning of “happening quickly” (see Figure 9).

(8) Ahmet gel-ebil-di. *Ahmet could come.* (ability)

(9) Yap-iver-di-m. *I did it quickly.* (action happening quickly)

Figure 9: Helping Verbs in Turkish

Much like the passive and causative forms, verbs attached with helping verbs have also been excluded from the corpus, for the same reasons. These helping verbs are highly productive, so their inclusion would lead to redundancy.

3.2. Selection of Entries to Be Included in the Corpus

In our analysis, we have made sure to include only the base forms of verbs, whether they are composed of a single word (ayamak “to sleep”) or a phrasal structure (rizya görmek “to see a dream or to dream”). What is meant by base forms? As discussed previously, derived forms that can be produced through a rule have not been included, as long as they share the same meaning with their bases. However, forms that have taken on different meanings, i.e. diverging from the bases from which they were derived, have been included. For instance, büyümek "to make it grow", is derived from the verb büyümek "to grow" through the addition of causative suffix “-t.” Hence, it is expected that büyümek means “make sb/sth grow”. Yet in time, it has gained a brand new meaning: to exaggerate, to overestimate. Thus, we included büyümek in our dataset even though it can be broken down to its components and produced from its base verb.

Another important concern is verbs that are actually base forms, despite being marked with passive or causative suffixes. While these verbs derive from actual roots, their verbal roots have either fallen out of use or simply cannot stand on their own. As such, they had to be included in the corpus; otherwise, there would be no entry they could derive from, and they could not be analyzed morphologically.

3.3. Exemplar and Further Explanation on TRopBank

The Turkish language makes extensive use of phrasal structures, metaphors and idioms. This reflects clearly on TRopBank, in comparison to the English PropBank. Instead of neatly arranged verbs consisting of a single word, TRopBank is filled with entries that are comprised of two, three or even more words. This issue would be of no consequence, were it not for the fact that said entries sometimes include the arguments as integrated parts of the verb. This leads to an interesting situation, where an argument cannot be properly annotated, as it is embedded within the entry...
itself. This creates the illusion that the entries lack said arguments. Yet the argument is simply an integrated part of the phrasal verb. *Hava bozmak* "(for the weather) to turn stormy, rainy or cloudy" is a representative instance of such verbs. The predicate, *hava bozmak* encapsulates its sole argument, *hava* "the weather", thus no argument is annotated for this verb.

For annotating the individual arguments, we have consistently focused on the definitions of the verbs, as some verbs take on different argument structures due to the differences in their meanings or usages.

4. The Annotation Process

4.1. Data Preparation

Before starting the annotation process, the first step was sifting through the data in the Turkish wordnet KeNet (Ehsani et al., 2018; Bakay et al., 2019a; Bakay et al., 2019b; Ozcelik et al., 2019) since the corpus had to be tidied up considerably. Many of the entries were either included accidentally, or were decided to be redundant. Certain nouns that were included in the list due to their morphological resemblance to verbs, such as *tokmak* "mallet", were excluded. Adjectival phrases were also excluded.

The second stage of the cleanup process was the removal of rule governed verbal derivations. As mentioned previously, these were mainly passive, causative and helping verb constructions. This stage presented a minor challenge: detecting a passive or causative suffix on the verb is not enough to remove it. The verb has to have a base form that can stand on its own and the base has to share its definition with the derived form. Verbs like *yürümek* "to walk" and *yürütmek* "to make sb/sth walk" fit this definition, thus *yürütmek* was removed from the data set.

As such, many entries had to be checked from the dictionary manually. Deciding whether an entry was a passive/causative structure that needed to be removed was not easy, and intuition had to be relied on in many cases.

After the redundant verbs were removed from the data set, verbs and their definitions were reviewed. Meanings of the verbs constituted the units, thus verbs were listed for each definition and merged if synonymous.

And finally, sample sentences were added for each entry in the data set. Some of these sample sentences were taken from a Turkish corpus, some were created by the annotators.

4.2. Main Issues Encountered During the Annotation Process

Once the data sorting process was finished, the task in hand was the annotations. However, this stage was the most time-consuming and it came with its own set of challenges. A wide array of non-obligatory bits of information have been included in the annotations in order to make sure that PropBank covers the entirety of the necessary information to process each verb. This presents the annotators with a difficult problem: to what extent should a piece of information be included as an argument of the verb? Subject and object(s) are always included as arguments as they are obligatory, thus, essential components of a verb.

We annotated each argument with the appropriate tag from our list of semantic roles. The tags used for marking semantic roles are as follows: (i) *PAG*: agent or experiencer, doer of action or experiencer of emotion; (ii) *PPT*: patient or theme, participant who is acted upon or undergoes change; (iii) *GOL*: goal or benefactive; goal of motion or recipient of action; (iv) *LOC*: location of event; (v) *DIR*: direction of motion; (vi) *SRC*: source of motion or event; (vii) *COM*: commitative, an instrument or a collaborating participant; (viii) *REC*: reciprocal, participant who reciprocates action; (ix) *TMP*: temporal, timing of event. We created cells from ARG0 to ARG4, and the maximum amount of arguments that a verb took was four (see Table 2 for examples).

Unlike the case of obligatory arguments, it was more challenging to decide whether to include information regarding the manner, time or place of the event. As such, the annotators have had to pay great attention to each entry, making sure to be consistent. General time and place information can be specified for any verb, therefore we chose not to include these as arguments. However, more specific occurrences of these have been included, such as "interval of time", or "place that relates to the structure of the event". For instance, in a sentence such as "I ate at a restaurant.", the place information is simply an additional detail and it is unrelated to the internal structure of the event. On the other hand, in "I went to the library.", the place information is an important component of the event, since "to go" is a verb of motion that entails a change in location. The same applies to temporal information. Only verbs that are inherently related to time were annotated with the TMP tag. Instruments, while not considered obligatory in theoretical syntax, have also been included in many instances.

Another challenge in the annotation process was to decide which verbs belong to the category of "unaccusatives". Being defined as a subcategory of intransitive verbs, which have only one argument, namely the subject; unaccusative verbs have only the subject argument, which is semantically the theme of the verb, i.e., it has the properties of an object despite occupying the subject position. Certain generalizations can be made about this type of verb: most of the time, they either express a change of state, or an inherent feature of its subject. Many verbs seem to be ambiguous when it comes to this categorization, and they seem to change category depending on context. Therefore, once again, intuition had to be relied upon for the classification of these verbs. How this manifests itself overtly in the annotations is that the verbs have an empty Arg0 slot (where we would normally expect the subject), and the subject is placed in the Arg1 slot (expected slot for objects). The verb "iňa olmak" ("to become prosperous") is an example for this.

Another point of interest is the presence of verbs that have zero arguments. These entries are few in number, and they occur mostly because all the available arguments are already embedded inside the phrasal verb. *İ ş başa düşmek* "to have to accomplish something on one’s own" can be considered as such idioms with zero arguments.

5. Statistics

For TRropBank, a total of 17,691 verbs were annotated. Around 1,000 verbs are to be added in the future, most of
which are idioms and verbs with zero arguments.

As the data suggests, unaccusative verbs that require a patient or theme in the ARG1 column constitute roughly 15.1% of all the annotated verbs (see Table 3). Based on the data, it can be inferred that Turkish has an evident preference for verbs that require an ARG0 over ones that require an ARG1 as their subject.

Moreover, we can see that a significant portion of Turkish verbs, 47.9% to be exact, have the transitive framework. Turkish displays an observable preference regarding transitivity.

Furthermore, having predicates that do not require any arguments, Turkish diverges from the majority of the languages whose PropBanks have been reviewed in Section 2. Even though predicates without arguments (idiomatic structures) make up less than 1% of the total, the existence of such a divergence is significant.

To sum up, TRopBank provides unprecedented data on the overall tendencies of Turkish verbs within the framework of transitivity and the portion of idiomatic expressions. As a result, we can infer that TRopBank helps us unveil the properties of argument structure of Turkish verbs in regards to theoretical linguistics in addition to being a valuable asset for NLP solutions.

### 6. Discussion

TRopBank, independently from its potential uses in NLP, shows a stark contrast between the verbal patterns of Turkish and English. Comparing the two corpora, one can observe that Turkish is very fond of phrasal structures and makes extensive use of idioms instead of simple verbs, as mentioned above. However, what is truly remarkable is the embedding of arguments inside the phrasal verb. How should these structures be analyzed? Here, we have opted to not include these arguments as separate annotations, but perhaps an alternate analysis could be implemented. The embedded arguments could be included in the annotations. Of course, this alternate account would come with its own complexities regarding how the analysis would be carried out. Turkish is structurally very different from Germanic or Romance languages, and this contrast needs to be properly accounted for in all future endeavors to process the language. Taking a model from English or other European languages and applying it to Turkish is not an easy task and hence, certain modifications need to be made in order for the analysis to succeed.

### Table 2: Examples from the Annotation Process

| ID            | SynSet          | Definition                                           | Example                                                                 | ARG0       | ARG1       | ARG2 | ARG3 |
|---------------|-----------------|-----------------------------------------------------|------------------------------------------------------------------------|------------|------------|------|------|
| TUR10-0902470 | içine ateş atmak | aşırı acı, sıkıntı veya üzüntü verecek davranışta bulunmak | Nazmiye’nin içine avuçla ateş atıp evden içeri giriyor ama başını kaldırp pencereye bakmıyordu. | acı veren kişi | verilen acı | acı verilen kişi |     |
| TUR10-0004750 | açıklamak       | Bir konuya ilgili gerekli bilgileri vermek, izah etmek | Hasan Şaq, bir soru üzerine, Güney Kore’de futbol oynamayı düşünmedikini açıkladı. | açıklama yapan kişi | açıkladığı şey | açıklama yapılan kişi |     |

### Table 3: Statistics from the Annotation Process

| Argument Type                      | Value | Percentage |
|------------------------------------|-------|------------|
| Verbs with no ARG0                 | 3023  | 17         |
| Verbs with no ARG1                 | 4486  | 25.3       |
| Verbs with no ARG2                 | 15803 | 89.3       |
| Verbs with no ARG0 but ARG1        | 2681  | 15.1       |
| ARG0                               | 14668 | 49.3       |
| ARG1                               | 13126 | 35.8       |
| ARG2                               | 1888  | 6.3        |
| ARG3                               | 78    | 0.26       |
| ARG4                               | 1     | 0.003      |
| pag                                | 14579 | 48.9       |
| ppt                                | 10665 | 44.1       |
| dir                                | 1431  | 4.8        |
| gol                                | 800   | 2.6        |
| loc                                | 814   | 2.7        |
| src                                | 604   | 2          |
| com                                | 481   | 1.6        |
| tmp                                | 156   | 0.5        |
| ext                                | 13    | 0.04       |
| Unaccusatives                      | 2681  | 15.1       |
| Verbs with no arguments            | 79    | 0.44       |
| Entries without a sample sentence  | 9941  | 56.1       |
| Intransitive verbs                 | 4180  | 23.5       |
| Transitive verbs                   | 8521  | 47.9       |
| Ditransitive verbs                 | 3043  | 17.2       |
| Total number of annotated entries  | 17691 |            |
| Total number of arguments          | 32755 |            |
| Average number of arguments        | 1.682 |            |

TRopBank is open to future improvements, especially regarding the level of detail in the annotations. The PAG and PPT tags can be further separated among themselves, with the addition of distinct tags for the roles of experiencer and patient. As a large-scale dataset, TRopBank has a great potential for augmenting the efficiency and accuracy of NLP applications within the framework of machine translation, information extraction and information retrieval (Ak et al., 2018). In addition, TRopBank provides a semantic information layer through the syntactic annotations. As a result, question-answering performance of NLP solutions gain a significant accuracy boost. Moreover, this semantic information layer can lead to more accurate and polished syntactic parsers.
Agirre, E. E., Aldezabal, I., Etxeberria, J., and Pociello, E. (2006). A Preliminary Study for Building the Basque PropBank. In Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC), April.

Ak, K., Yildiz, O. T., Esgel, V., and Toprak, C. (2018). Construction of a Turkish proposition bank. Turkish Journal of Electrical Engineering and Computer Science, 26:570 – 581.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., and Estarrona, A. (2010a). Building the Basque propbank. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta, may. European Language Resources Association (ELRA).

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., Estarrona, A., and Uria, L. (2010b). Euspropbank: Integrating semantic information in the basque dependency treebank. In Computational Linguistics and Intelligent Text Processing, pages 60–73, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anwar, M., Bhat, R. A., Sharma, D., Vaidya, A., Palmer, M., and Khan, T. A. (2016). A proposition bank of Urdu. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may. European Language Resources Association (ELRA).

Bakay, O., Ergelen, O., and Yildiz, O. T. (2019a). Integrating Turkish WordNet KeNet to Princeton WordNet: The case of one-to-many correspondences. In Innovations in Intelligent Systems and Applications.

Bakay, O., Ergelen, O., and Yildiz, O. T. (2019b). Problems caused by semantic drift in wordnet synset construction. In International Conference on Computer Science and Engineering.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). Propbank: Semantics of new predicate types. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), European Language Resources Association (ELRA).

Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., and Graça, J. (2012). A propbank for Portuguese: the cintil-propbank. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey, may. European Language Resources Association (ELRA).

Burchardt, A., Frank, A., Pado, S., and Pinkal, M. (2006). The salsa corpus: a German corpus resource for lexical semantics. In Proceedings of LREC 2006 : the 5 th International Conference on Language Resources and Evaluation, Genoa, Italy. - Paris, pages 969–974.

Şahin, G. G. and Adali, E. (2017). Annotation of semantic roles for the Turkish proposition bank. Language Resources and Evaluation, May.

Şahin, G. G. (2016a). Framing of verbs for Turkish propbank. In TurCLing 2016 in conj. with 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016).

Şahin, G. G. (2016b). Verb sense annotation for Turkish propbank via crowdsourcing. In 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2016).

De Clercq, O., Hoste, V., and Monachesi, P. (2012). Evaluating automatic cross-domain dutch semantic role annotation. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, pages 88–93.

Ehsani, R., Solak, E., and Yildiz, O. (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.

Erk, K., Kowalski, A., Padó, S., and Pinkal, M. (2003). Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. ACL ’03, pages 537–544, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fillmore, C. J., Ruppenhofer, J., and Baker, Collin, F. (2004). FrameNet and Representing the Link between Semantic and Syntactic Relations, pages 19–62. Language and Linguistics Monographs Series B. Institute of Linguistics, Academia Sinica, Taipei.

Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternations. In Causatives and transitivity, Studies in Language Companion Series B. Institute of Linguistics, Academia Sinica, Taipei.

Haverinen, K., Kanerva, J., Kohonen, S., Missilli, A., Ojala, S., Viljanen, T., Laippala, V., and Ginter, F. (2015). The Finnish proposition bank. Language Resources and Evaluation, 49(4):907–926, Dec.

Kawahara, D., Kurohashi, S., and Hasida, K. (2002). Construction of a Japanese relevance-tagged corpus. In LREC. European Language Resources Association.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In LREC. European Language Resources Association.

Kingsbury, P. and Palmer, M. (2003). Propbank: The next level of treebank. In Proceedings of Treebanks and Lexical Theories, Växjö, Sweden.

Kipper, K., Dang, H. T., and Palmer, M. (2000). Class-based construction of a verb lexicon. In Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 691–696. AAAI Press.

Levin, B. (1993). English verb classes and alternations : a preliminary investigation. of Chicago Press, University.

Márquez, L., Villarejo, L., Martí, M. A., and Taulé, M. (2007). Semeval-2007 task 09: Multilevel semantic annotation of Catalan and Spanish. In Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval ’07, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mirzaei, A. and Moloodi, A. (2016). Persian proposition bank. In Proceedings of the Tenth International Con
ference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.

Mukund, S., Ghosh, D., and Srihari, R. K. (2010). Using cross-lingual projections to generate semantic role labeled corpus for Urdu: A resource poor language. In Proceedings of the 23rd International Conference on Computational Linguistics, COLING ’10, pages 797–805, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ozcelik, R., Parlar, S., Bakay, O., Ergelen, O., and Yildiz, O. T. (2019). User interface for Turkish word network KeNet. In Signal Processing and Communication Applications Conference.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. Comput. Linguist., 31(1):71–106, March.

Palmer, M., Ryu, S., Choi, J., Yoon, S., and Jeon, Y. (2006). Korean propbank. Philadelphia: Linguistic Data Consortium.

Palmer, M., Babko-Malaya, O., Bies, A., Diab, M., Maamouri, M., Mansouri, A., and Zaghouani, W. (2008). A pilot Arabic propbank. In The Sixth International Language Resources and Evaluation Conference (LREC2008), May.

Song, H., Park, C., Lee, J., Lee, M., Lee, Y., Kim, J., and Kim, Y. (2012). Construction of korean semantic annotated corpus. In Computer Applications for Database, Education, and Ubiquitous Computing - International Conferences, ELMUNET and ENESST 2012, Held as Part of the Future Generation Information Technology Conference, (FGIT) 2012, Gangneug, Korea, December 16-19, 2012. Proceedings, pages 265–271.

Sulubacak, U., Pamay, T., and Eryiğit, G. (2016). Imst: A revisited Turkish dependency treebank. In The First International Conference on Turkic Computational Linguistics, pages 1–6.

Vaidya, A., Choi, J. D., Palmer, M., and Narasimhan, B. (2011). Analysis of the Hindi proposition bank using dependency structure. In Proceedings of the 5th Linguistic Annotation Workshop, LAW V ’11, pages 21–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

van der Plas, L., Samardžić, T., and Merlo, P. (2010). Cross-lingual validity of propbank in the manual annotation of French. In Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV ’10, pages 113–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

X., N. and Palmer, M. (2009). Adding semantic roles to the Chinese treebank. Natural Language Engineering, 15(1):143–172.

Xue, N. (2006). A Chinese semantic lexicon of senses and roles. Language Resources and Evaluation, 40(3):395–403, Dec.

Yıldız, O. T., Solak, E., Çandır, Ş., Ehsani, R., and Görgün, O. (2015). Constructing a Turkish constituency parse treebank. In Information Sciences and Systems 2015, pages 339–347. Springer.

Zaghouani, W., Diab, M., Mansouri, A., Pradhan, S., and Palmer, M. (2010). The revised Arabic propbank. In Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV ’10, pages 222–226, Stroudsburg, PA, USA. Association for Computational Linguistics.