

# Persian Proposition Bank

Azadeh Mirzaei\*, Amirsaeid Moloodi\*\*

\*Department of Linguistics, Allameh Tabataba'i University, Tehran, Iran \*\*Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran  
E-mail: azadeh.mirzaei@atu.ac.ir, amirsaeid.moloodi@shirazu.ac.ir

## Abstract

This paper describes the procedure of semantic role labeling and the development of the first manually annotated Persian Proposition Bank (PerPB) which added a layer of predicate-argument information to the syntactic structures of Persian Dependency Treebank (known as PerDT). Through the process of annotating, the annotators could see the syntactic information of all the sentences and so they annotated 29982 sentences with more than 9200 unique verbs. In the annotation procedure, the direct syntactic dependents of the verbs were the first candidates for being annotated. So we did not annotate the other indirect dependents unless their phrasal heads were propositional and had their own arguments or adjuncts. Hence besides the semantic role labeling of verbs, the argument structure of 1300 unique propositional nouns and 300 unique propositional adjectives were annotated in the sentences, too. The accuracy of annotation process was measured by double annotation of the data at two separate stages and finally the data was prepared in the CoNLL dependency format.

**Keywords:** semantic role labeling, Persian Proposition Bank, predicate-argument information, propositional noun, propositional adjective, CoNLL dependency format.

## 1. Introduction

As we know, the amount of information obtained from the syntactic analyses of sentences is limited and one needs to go to a higher level which is semantic analyses to achieve the propositional content of the sentence. The following examples elaborate on the necessity of semantic role labeling.

1. Ali broke the window.  
OBJ
2. The window broke.  
SBJ

In the above sentences, the active verb "to break" has two different syntactic realizations. In the first sentence "the window" is the verb's direct object and its subject in the second one. Although "the window" has two different syntactic roles in these sentences, it plays the same underlying semantic role (PATIENT) in both. Conversely, in some cases the same syntactic realizations may result in different meanings.

3. ?ali be so?al barmigardad  
Ali to question gets back  
'Ali gets back to the question.'

ARG0: ?ali (Ali)  
ARG1: be so?al (to the question)  
REL: barmigardad (gets back)

4. mozu? be bamerizi barmigardad  
issue to planning gets back  
'The issue gets back to planning.'

ARG1: mozu? (issue)  
ARG2: be bamerizi (to planning)  
REL: barmigardad (gets back)

As it is obvious from the examples (3) and (4), the syntactic realizations of the both sentences are the same (||subject, prepositional object (to)||) but the assigned numbered arguments are different and so there are semantically two different verbs. As a result the semantic analysis of the sentences is essential to achieve propositional content.

The first Persian Proposition Bank like the other Proposition Banks (Kingsbury & Palmer, 2002; Palmer et al., 2005; Burchardt et al., 2006; Xue & Palmer, 2009; Palmer et al., 2008; Ohara et al., 2004; Taulé et al., 2008) aims to construct the argument structure of the predicates and in order to analyze the sentences semantically, the functional tags are specified too. All of these Proposition Banks which are mainly constructed according to the verb classification of Levin (1993) and Dang et al. (2000), serve as the basis of the machine learning classifiers in the natural language processing tasks such as Surdeanu et al. (2008), Hajič et al. (2009) and Gildea & Jurafsky (2002).

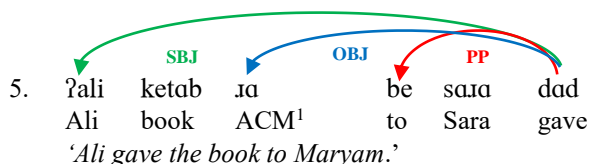
## 2. Persian Proposition Bank

Persian Proposition Bank added a layer of predicate-argument information to the syntactic structures of PerDT (Rasooli et al. 2013), a manually syntactically annotated Persian corpus. The data was 29982 sentences which were syntactically annotated according to dependency grammar in which each word had one head and the head of the sentence, often the verb, was the dependent of an artificial object called the root word (Kübler et al. 2009).

In PerPB the first candidates for semantic role labelling were the dependents of the verbs which had been specified in PerDT. It is notable to mention that our semantic approach to define a verb was different from the syntactic approach and so there may be some changes in argument selection and annotation. In the other words, each event or action regardless of its syntactic form was treated as verb in PerPB. For example in PerDT there is a metaverb

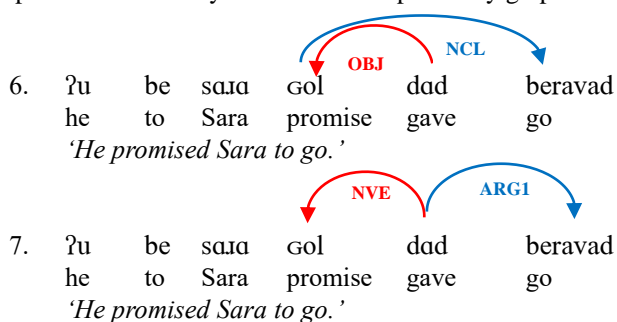
"dadan" to give which has a meta syntactic structure as shown below:

||subject, object, prepositional object (to)||



In the above sentence, the word "dad" as the head of the sentence has three dependents which have been specified with three different colors and three different syntactic labels. The direction of the dependency graph is from the head to the (head of) dependent. For example, "be" to as the head of the prepositional phrase "be marjam" to Maryam is considered as one of the dependents of "dad".

According to the syntactic approach of PerDT, some word sequences like "gol dadan" (promise + to give) to promise and "?edzaze dadan" (allow + to give) to allow have been considered to have the above-mentioned structure in which the words "gol" and "?edzaze" are the objects of the simple verb "dadan" not the non-verbal elements of the complex predicate "gol dadan" or "?edzaze dadan"; more clearly the above word sequences are not considered as complex predicates and considered just as a combination of a noun (with the syntactic role object) and a simple verb. However since in PerPB every event or action with inflectional properties of the verbs in the Persian is treated as the verb, the previously mentioned sequences are considered individual verbs. In the other words, many of the verb complements in PerDT now have become non-verbal element of a new complex predicate in PerPB. A non-verbal element is the first part of a compound verb or complex predicate which is a noun, an adjective or a preposition; for example, in the Persian compound verb "?e:rsal kardan" (sending + to do) to send, "?e:rsal" is considered as a non-verbal element. These changes in the approach of the verb or complex predicate definition resulted in an increase in the number of the verbs from more than 4500 to more than 9200 unique verbs while the sentences in the syntactic corpus did not change. The mentioned differences of the two approaches are represented below by the means of dependency graphs:



Sentence (6) shows the syntactic approach in which to the word "gol" promise the label *object* with OBJ abbreviation has been assigned and the clause "be.avad" with the label

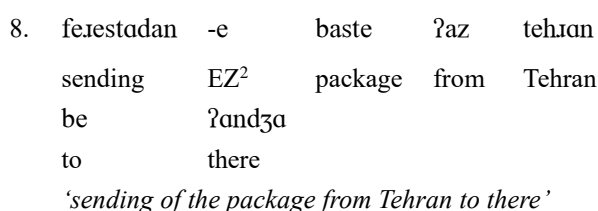
*clause of noun* (NCL) is the dependent of "gol". Therefore in the syntactic approach, "gol dadan" is not a complex predicate and it is a combination of an object and a simple verb. However the semantic approach of PerPB (reflected in sentence 7) considers the word "gol" as the non-verbal element (NVE) of the complex predicate "gol dadan" to promise and thus the clause "beravad" is considered as a semantic argument (ARG1) of the complex predicate.

It should be noted that the present project is not only limited to semantic role labeling of the verbal predicates but all the dependents of nominal and adjectival predicates are annotated too (something like Bonial et al., 2014) and so this project is a combination of Persian ProBank and NomBank.

For the head of a noun phrase to be considered a markable noun it has to satisfy the following two criteria:

1. It should be a propositional noun and have at least one visible argument in the sentence/ phrase.
2. It should be a propositional noun and have at least one visible adjunct in the sentence/ phrase.

Propositional nouns are defined as nouns which have a propositional role or argument structure like verbs, however they do not assign nominative or accusative case to their arguments and do not have the verb inflection. As an example, consider the noun "dismissal" in "dismissal of the employees from the company". Since the word has argument structure, it belongs to propositional nouns. The two arguments of the noun in the mentioned sentence are ARG1 (the employees) and ARG2 (from the company). It is noteworthy that the basis of noun semantic role labeling is the verb semantic roles. In other words, all the nouns that derive semantically or morphologically from the related verbs are labeled exactly like the verbs. For example, the noun "sending" is labeled like the verb "to send":



ARG1: baste (package)  
ARG3: ?az teh:an (from Tehran)  
ARG4: be ?andza (to there)  
REL: fe:restadan (sending)

Like propositional nouns, propositional adjectives are defined as adjectives which have argument structure. The two mentioned criteria for a nominal head in order to be considered as markable are also the same for the Persian adjectives. It should be noted that since almost all the Persian adjectives in predicative function are considered non-verbal elements of complex predicates, only the dependents of attributive adjectives are semantically

<sup>1</sup> Accusative case marker

<sup>2</sup> Ezafe dependent

annotated. As an example consider the phrase "manabe?-e modʒud dar ʔi:ɾɒn" *available resources in Iran*. Since the attributive adjective "modʒud" *available* has argument structure it belongs to propositional adjectives. The two arguments of the adjective are ARG1 (manabe? *resources*) and ARG2 (dar ʔi:ɾɒn *in Iran*).

### 3. Semantic Roles

The basis of the names and definitions of the semantic roles in the PerPB is the work of Fillmore (1976), Fillmore & Atkins (1998) and Fillmore & Baker (2001). Additionally, the basis of the (verbal, nominal and adjectival) predicate classification of the PerPB is the verb classification of Levin (1993) and Verbnets (Kipper et al., 2006). The 26 semantic roles (AGENT, PATIENT, THEME, EXPERIENCER, etc.) of the PerPB have been grouped together to form the numbered arguments in the way of PropBank (Kingsbury & Palmer, 2002). So there exists two different kinds of labels. The first is the numbered arguments that begins from 0 and ends at 5 and each of the numbered arguments corresponds to some specific semantic roles while the second one is the labels denoting adverbial elements and some discourse labels (ARGM).

ARG0 is generally an argument that Dowty (1991) called prototypical AGENT and includes the participants that volitionally involve in the event or state, cause something to happen or have movement relative to the position of another participant. In PerPB, ARG0 corresponds to four semantic roles of AGENT, CAUSE, STIMULUS and EXPERIENCER. In Dowty's view, the semantic roles PATIENT and THEME which are generally affected by another participant or are stationary relative to the movement of another participant are considered as the prototypes of ARG1. ARG2 generally plays the ATTRIBUTE, RECIPIENT role or the starting point of an action when the ARG1 has been already present in the sentence. ARG3 and ARG4 are mainly related to the verbs of motion. ARG3 is usually the starting point of an action or event while ARG4 is usually the ending point of an action or event. ARG5 is the other numbered argument in PropBank but as the incorporation of objects with the verbs is a usual phenomenon in Persian, the maximum number of arguments for a given verb is five and ARG5 does not exist in PerPB. In addition to numbered arguments, there is another argument called ARG6 which is the CAUSE of action verbs. Accordingly for the verbs with two animate AGENTs, the CAUSE is annotated as ARG6 and the AGENT as ARG0.

9. Ali walked his dog.  
ARGA: Ali  
ARG0: his dog  
REL: walk

The second kind of labels in PerDT are either adverb or adjunct. These optional and non-required elements of PerDT including circumstantial, mood and textual adjuncts are annotated as the functional tags in PerPB. Table 1 shows the list of these adjuncts.

ADV:	Adverbials
CAU:	Cause
COM:	Comitative
CON:	Conditional
DIR:	Directional
DIS:	Discourse
EXT:	Extent
GOL:	Goal
INS:	Instrument
LOC:	Locative
MNR:	Manner
MOD:	Modal
NEG:	Negation
PRD:	Secondary Predication
PRP:	Purpose
RCL:	Relative Clause Link
REC:	Reciprocals
RPT:	Repetition
TMP:	Temporal
TOP:	Topicality

Table 1: Functional tags in Persian PropBank

Circumstantial adjuncts represent additional information about TIME, LOCATION, CAUSE, PURPOSE, DIRECTION, MANNER, EXTENT, REPETITION and so on. Comment and mood adjuncts show the view of the speaker or the writer about the proposition or proposal, like modal verbs that indicate the degree of either the possibility or probability of the verb realization. Finally textual adjuncts function as the connectors between the clauses or sentences like discourse adjuncts.

We have three other types of adverbs which are related to the participants (the numbered arguments) and so they are different from the other functional tags. These adverbs are COMITATIVES, RECIPROCALs and SECONDARY PREDICATIONs. COMITATIVE modifier indicates either a natural person or a legal one who accompanies the participant of the main event (typically ARG0):

10. ʔali ba dust -aʃ be pa:k  
Ali with friend his to park  
raft  
went  
'Ali went to the park with his friend.'

ARG0: ʔali (Ali)  
ARGM-COM: ba dustaʃ (with his friend)  
ARG4: be pa:k (to the park)  
REL: :raft (went)

RECIPROCALs which include reflexives and reciprocal pronouns (e.g. himself, itself, themselves and each other) usually have overt antecedences annotated as the numbered arguments in the sentences but in the pro-drop Persian language it is possible that some of the reflexives lose their antecedent by the deletion of the subject pronoun in the

sentence. Even in this case, they are annotated as RECIPROCAL not as numbered argument and the antecedent is inferable by the subject-verb agreement:

11. xodaʃ be daneʃgah ʃaft  
 himself to university went  
 \*Himself<sup>3</sup> went to the university.

ARGM-REC: xodaʃ (himself)  
 ARG4: be daneʃgah (to the university)  
 REL: ʃaft (went)

Sentence (13) is an example in which the reflexive "himself" has been accompanied with its antecedent (Ali) but in the above sentence, the antecedent of the reflexive pronoun "himself" has been omitted and the sentence is still grammatical in Persian (not in English). Even in this case that the reflexive plays the syntactic role of subject, it is annotated as RECIPROCAL in PerPB.

SECONDARY PREDICATION is an adjunct that modifies an argument of a verb instead of the verb itself or the whole sentence. This kind of adjunct semantically plays the role of attribute for the argument which it modifies.

12. ʔali segeftzade vaʔd -e ʔotaʒ ʃod  
 Ali surprised enter EZ room did  
 '(While) surprised, Ali entered the room.'

ARG0: ʔali (Ali)  
 ARGM-PRD: segeftzade (surprised)  
 ARG4: ʔotaʒ (room)  
 REL: vaʔed ʃod (entered)

The following example shows the assignment of some of the mentioned roles in a Persian sentence.

13. ʔali xodaʃ di.ruz kado ʃa  
 Ali himself yesterday gift ACM  
 ʔaz fo.ruʃgah ba.rʃ -e  
 from market for EZ  
 madar -aʃ xarid  
 mother his bought

'Ali bought the gift from the market for his mother yesterday.'


ARG0: ʔali (Ali)  
 ARGM-REC: xodaʃ (himself)  
 ARGM-TMP: di.ruz (yesterday)  
 ARG1: kado ʃa (the gift)  
 ARG2: ʔaz fo.ruʃgah (from the market)  
 ARGM-GOL: ba.rʃ-e mada.r-aʃ (for his mother)  
 REL: xa.ʔid (bought)

Based on their semantic functions, GOAL and INSTRUMENT modifiers are very similar to the arguments; however the difference between them and the arguments is that they do not belong to the argument

structure of the verbs. GOAL adjunct like a BENEFICIARY argument is an animate endpoint of a transferred item but like the other adjuncts it is not in the argument structure of the verb. INSTRUMENT is a medium whereby the action of a verb is accomplished.

Now in this section we want to briefly compare the functional tags of PerPB with PropBank. Unlike PropBank, adverbs of frequency (eg. baʔzi oʒat *sometimes*, hamiʃe *always*, hargez *never*) and repetition (dobare *again*) belongs to the functional tag REPETITION While in PropBank these concepts belong to the label TEMPORAL with the exception of *never* which is considered as a member of NEGATION label. Unlike many languages like German, English, Italian, Russian, etc., in Persian the negation particle is not an autonomous constituent inserted before or after a verb phrase but it is a prefix which attaches to the verb; so a morphological analysis would be needed prior to annotation and negation can't be annotated again in this level and the only NEGATION label in PerPB is just used for "na.... na...." *not.... not....* constructions. In PropBank, conditionals are the members of the ADVERBIAL functional tag, however they constitute the CONDITIONAL label in PerPB.

Persian as an analytic language represents some different relational meanings and grammatical categories by using two or more words rather than inflection. Accordingly in analytic form of operators, the label OP is used to show the fragmented operator of the verbs.

14. sara da.r hale neveʃtan ʔast  
 Sara in situation to write is  
 'Sara is writing'
- 

In (14) the progressive mood has been encoded by more than one word ("da.r hale X budan"); so the OP label is used.

MOD is a more comprehensive concept in PerPB in comparison with Propbank. MODs are modal verbs in PropBank but in PerPB they are any realization of the epistemic or evidential concepts whether as a modal verb, an adverb, dependent clause, prepositional phrase, etc. So in PerPB *perhaps*, *may*, *it's possible*, etc. are MOD.

Finally, it should be mentioned that Persian is a free word order language and most of the syntactic movements are applied to express some discourse functions. In PerPB these meaningful movements are specified with two labels, TOPICALITY and RELATIVE CLAUSE LINK. The first label is used to annotate the topicalized constituents and the second one deals with the distance between a relative clause and its head. It means that if the relative clause is strictly adjacent to its head, the label RCL0 is assigned to it; however in cases where at least one constituent appears between the relative clause and its head, the relative clause receives the label RCL1. TOP and RCL labels are so useful to clearly show the relationship between the separated parts of an argument. The three following examples represent how these labels are assigned.

<sup>3</sup> The asterisk (\*) denotes the ungrammaticality of a structure.

15. ?ali .ia dust -a? .ia didam  
 Ali ACM friend his ACM saw  
 'Ali, his friends I saw'

TOP

16. fard -i ke mixandad xo?hal ?ast  
 one IM<sup>4</sup> that smiles happy is  
 'the one who smiles is happy'

RCL0

17. fard -i xo?hal ?ast ke mixandad  
 one IM happy is that smiles  
 'the one who is happy smiles'

RCL1

#### 4. Annotation Procedure

Since by the time of PerPB project there were not any Persian semantic datasets neither in the form of PropBank nor in the form of VerbNet, a combination of PropBank and VerbNet approach was applied in PerPB semantic role annotation which resulted in one of the most important by products of the project named as Persian Semantic Valency Lexicon (Mirzaei & Moloodi, 2015); In the annotation procedure, propositional heads were annotated based on both numbered approach of PropBank and semantic roles approach of VerbNet. So a detailed and comprehensive guideline including the accurate definition of semantic roles and their systematic correspondences to numbered arguments and abundant examples to elaborate them was prepared. Through the process of semantic role labeling, the syntactic description of the sentences was available to the annotators. The process of semantic role labelling by the annotators can be stated like this: At first, the annotators chose the verbs or any other propositional heads having at least an argument or adjunct and then the direct syntactic dependents of all verbs or propositional heads were selected as the first candidates for being annotated. The third stage of semantic role annotating was the assigning of a label to the selected dependents based on their semantic function in the sentences or phrases. In this stage, all the arguments were annotated both by numbered approach and semantic roles approach and a considerable attention was paid by the annotators to correctly annotate all the non-verbal element dependents in PerDT as one of the verb arguments in PerPB

Considering the annotation procedure, there is a significant difference between PerPB and PropBank. The annotation procedure of latter one began by defining the Frame Files first. This was also true for the tectogrammatic tagging of the Prague Dependency Treebank (Hajič, 2005) in which at first a valency lexicon was prepared and then it was used as a guide to annotate sentences. However, in PerPB there were no Frame Files prior to the building of the corpus and the construction of the valency lexicon was done

<sup>4</sup> Indefinite marker

simultaneously as the corpus annotation was proceeded.

It also has to be noted that some sentences were annotated twice by two different annotators at two different phases, one at the beginning of the annotating process (about 4000 sentences) and the other at the end (about 2000 sentences). If the annotations were different, they were reported to two supervisors (the authors of the present article) who would select the correct one or if they were both incorrect, they would introduce the correct annotation. It is noteworthy that the double annotation phase was completely separate from the correction phase; in the other words, for measuring the inter-annotator agreement (introduced in section 6) the double annotated sentences were preserved as a separate version of the corpus.

#### 5. Annotators

The annotators consisted of four PhD candidates (linguistics), and five MA graduates (three linguistics graduates, one Persian language and literature graduate and one computational linguistics graduate) who were native Persian speakers. Annotators were presented and trained with a comprehensive guideline, describing all the semantic roles with abundant examples.

#### 6. Annotators Agreement

Agreement between the two annotators is measured using the kappa statistics (Cohen, 1960), which is defined with respect to the probability of inter-annotator agreement,  $P(A)$ , and the agreement expected by chance,  $P(E)$ :

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

The measurements were performed in two individual phases, once at the beginning and once at the end of the work.

		K (Arg)	K (ArgM)
role classification	first	0.899057	0.863363
	second	0.910157	0.904402

		K
role identification	First	0.939316
	Second	0.954103

Table 2: Inter-annotator agreement

Considering any direct dependent of the verb and non-verbal element as a candidate for semantic role labeling led the agreement to be much higher than the chance and this made us measure the statistics of role classification and role identification separately. The Kappa statistics for role identification and classification are shown in Table 2. As the table shows, agreement on the role identification and the role classification is very high in two stages.

## 7. Statistics

Finally 29982 sentences were manually annotated. The total number of annotated distinct verbs was about 9200 and the total number of annotated nouns and adjectives was about 1300 and 300 respectively.

Number of Sentences	29982
Average Sentence Length	16.61
Number of Verbs	62889
Number of Verb Lemmas	>9200
Number of distinct propositional Nouns	1300
Number of distinct propositional Adjectives	300

Table 3: Statistics about the frequency of words in the PerPB

<b>ARG1</b>	27.2%
<b>ARGM</b>	24.2%
<b>NVE (semantic)<sup>5</sup></b>	22.1%
<b>ARG0</b>	11.9%
<b>ARG2</b>	7.4%
<b>ARG4</b>	0.71%
<b>ARG3</b>	0.24%
<b>ARGA</b>	0.07%

Table 4: The eight most frequent semantic roles

Tables 4, 5, 6 and 7 show statistics for the corpus. Table 4 shows the eight most frequent semantic roles of the corpus

<b>ARG1</b>	OBJ +VCL	37.2	SBJ	20.1	MOZ	14.1	NPP	15.4	VPP	5.1
<b>ARGM</b>	ADV+AJUCL	69	NPOSTMOD	9.4	NPREMOMD	7.1	ROOT	3.5	MOZ	2.9
<b>NVE (semantic)</b>	NVE (syntactic)	66.8	MOS	13.3	OBJ	6.5	VPP	3.4	VPRT	2.2
<b>ARG0</b>	SBJ	80.2	MOZ	10.7	POSDEP	2.1	ADV	1.3	POSDEP	1.2
<b>ARG2</b>	MOS	26.4	NPP	22.6	VPP	18.2	ADV	6.1	OBJ+VCL	5.3

Table 6: The five most frequent semantic labels and their corresponding syntactic roles (percentages)

<b>SBJ</b>	ARG0	60.5	ARG1	34.6	NVE (semantic)	1.1	ARG2	<1	ADV	<1
<b>NVE (syntactic)</b>	NVE (semantic)	99.5	ARG1	<1	ARG0	<1	ARG2	<1	ADV	<1
<b>OBJ</b>	ARG1	79.5	NVE (semantic)	14.8	ARG0	1.5	ARG2	1.1	ADV	<1
<b>MOS</b>	NVE (semantic)	56.5	ARG2	37.5	ARG1	2.6	ARG0	1.9	ADV	<1
<b>VPP</b>	ARG1	34.1	ARG2	33	NVE	18.6	ARG4	8	ARG3	2.1

Table 7: The five most frequent syntactic labels and their corresponding semantic roles (percentages)

including numbered arguments, functional roles and semantic NVEs. Table 5 shows the ten most frequent functional roles of the corpus. Table 6 shows the five most frequent semantic labels associated with various syntactic positions. Table 7 shows the five most frequent syntactic labels, headed by verb, associated with various counterpart semantic labels.

<b>Ext</b>	18.4%
<b>TMP</b>	15.3%
<b>ADV</b>	10.5%
<b>MNR</b>	10.1%
<b>LOC</b>	8.1%
<b>PRD</b>	7.5%
<b>RPT</b>	4.8%
<b>CON</b>	4.01
<b>PRP</b>	4%
<b>CAU</b>	3.3%

Table 5: The ten most frequent functional roles

To simplify reading of table 6, one of its five rows is explained. For example, the fifth row shows that the semantic role ARG2 has five syntactic counterparts with different percentages of occurrence in the corpus; in other words, 26.4 percent of the semantic role ARG2 had the syntactic label *mosnad*<sup>6</sup> (MOS), 22.6 percent of ARG2s had the syntactic label *preposition of noun* (NPP) and so forth.

<sup>5</sup> The *semantic* or *syntactic* adjective beside the NVE (non-verbal element) label in the tables represents the two afore-mentioned (section 2) different approaches in the complex predicate

definition.  
<sup>6</sup> attribute

As the statistics shows, like PropBank Arg1 is the most frequent label among different semantic roles and the most frequent syntactic counterpart of Arg1 is object. As the table 4 shows, among the numbered arguments, in PerPB like PropBank, Arg1, Arg0 and Arg2 respectively are the most frequent arguments.

## 8. Conclusion

This paper presented the procedure of the development of the first Persian Proposition Bank, which added semantic role labels to PerDT. In order to achieve the highest accuracy, the annotators were presented with a detailed and comprehensive guideline and several guide notes and their annotations were checked continuously by the supervisors. In purpose to evaluate the accuracy of annotation, we calculated the inter-annotator agreement that was assessed as very high in two different stages.

Besides the semantic role labeling of verbs in the present corpus, the dependents of propositional nouns and adjectives were semantically annotated too.

As mentioned in the annotation procedure section, the data was prepared based on a combination of VerbNet and PropBank approach. The different approaches led to two separate data packs. The first one like PropBank comprises of the numbered arguments and the functional tags of propositional heads, while the second one like VerbNet consists of specific semantic roles and the functional tags of propositional heads. For more information about the release of the first version of PerPB we introduce the website <http://www.peykaregan.com>.

## 9. Acknowledgements

The project was funded by Computer Research Center of Islamic Sciences (CRCIS). We really appreciate the linguists who helped us in annotating: Parinaz Dadras, Saeedeh Ghadrdoost-Nakhchi, Manoucher Kouhestani, Mostafa Mahdavi, Neda Poormorteza-Khameneh, Morteza Rezaei-Sharifabadi, Fatemeh Shafie and Salimeh Zamani; and the programmers who helped us in the process of the development of the corpus: Hooman Mahyar and Fatemeh Sedghi; and other colleagues especially Mahdi Behniafar.

## 10. References

- Bonial, C., Bonn, J., Conger, K., Hwang, J. D. and Palmer, M. (2014). PropBank: Semantics of New PredicateTypes. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, pp. 3013-3019.
- Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S. and Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, pp. 969-974.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychosocial measurement*, 20(1), pp. 37-46.
- Dang, H. T., Kipper, K. and Palmer, M. (2000) Integrating compositional semantics into a verb lexicon. In *Proceedings of the Eighteenth International Conference on Computational Linguistics*. Saarbrücken, Germany, pp. 1011-1015.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3), pp. 547-619.
- Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1), pp. 20-32.
- Fillmore, C. J. and Atkins, B. S. (1998). FrameNet and lexicographic relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998)*. Granada, Spain, pp. 417-423.
- Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of NAACL WordNet and Other Lexical Resources Workshop*. Pittsburgh, USA.
- Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3), pp. 245-288.
- Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank. In M. Šimková (Ed.), *Insight into Slovak and Czech Corpus Linguistics*. Bratislava, Veda, pp. 54-73.
- Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J. and Straňák, P. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*. Boulder, USA, pp. 1-18.
- Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain, pp. 1989-1993.
- Kipper, K., Korhonen, A., Ryant, N. and Palmer, M. (2006). Extending VerbNet with novel verb classes. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy, pp. 1027-1032.
- Kübler, S., McDonald, R. and Nivre, J. (2009). *Dependency parsing. Synthesis Lectures on Human Language Technologies*. Morgan and Claypool Publishers.
- Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago press.
- Mirzaei, A. and Moloodi, A. (2015). The PropBank-Based Persian Semantic Valency Lexicon. In *Proceedings of International Workshop on Treebanks and Linguistic Theories (TLT14)*. Warsaw, Poland, pp. 284-291.
- Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H. and Ishizaki, S. (2004). The japanese framenet project: An introduction. In *Proceedings of the Satellite Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Fourth international conference on Language Resources and Evaluation (LREC 2004).

- Lisbon, Portugal, pp. 9-11.
- Palmer, M., Babko-Malaya, O., Bies, A., Diab, M. T., Maamouri, M., Mansouri, A. and Zaghouni, W. (2008). A Pilot Arabic Propbank. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, pp. 3467-3472.
- Palmer, M., Gildea, D. and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1), pp. 71-106.
- Rasooli, M. S., Kouhestani, M. and Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, USA, pp. 306-314.
- Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L. and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Manchester, UK, pp. 159-177.
- Taulé, M., Martí, M. A. and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco, pp. 96-101.
- Xue, N. and Palmer, M. (2009). Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1), pp. 143-172.