5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 227-231, Poznań, Poland, 2011.

A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, Behrouz Minaei-Bidgoli

Computer Engineering Department, Iran University of Science and Technology, Tehran, Iran
{rasooli@comp.,b_minaei@}iust.ac.ir
Department of Linguistics, University of Tehran, Tehran, Iran
{m.kouhestani@alumni.,a.moloodi@}ut.ac.ir

Abstract

Valency lexicons are valuable resources for natural language processing. The need for new resources for languages encourages researchers to collect new datasets. One of the most important datasets is valency lexicons. In valency lexicons, information about obligatory and optional complements of words is annotated at the syntactic and semantic levels. In this paper, we report the development of the first syntactic valency lexicon of Persian verbs. This lexicon is part of the Persian Dependency Treebank Project. The lexicon consists of 4282 distinct verb lemmas and 5429 distinct verb-valency pairs.

Keywords: Valency lexicon, Syntactic valency, Persian, Dependency treebank.

1. Introduction[1]

Lexicons and treebanks are considered valuable resources in computational linguistics and natural language processing. One of these main resources is valency lexicons. This kind of resource is very important for free-word order languages like Czech (Hlaváčková, Horák, and Kadlec, 2006). In valency lexicons, there exists information about obligatory and optional complements of words (mostly verbs and occasionally nouns and adjectives). The notion of valency originates from dependency grammar (henceforth DG). DG is a syntactic theory in which syntactic structure is determined by the relation between a head and its dependents.

In the Persian language, the abundance of light verb constructions (LVCs) increases the need for a feasible list of verbs. The need stems from the fact that it is much easier for the computer to identify a simple verb than a compound one. When confronted by a sentence containing a compound verb, the machine has to decide which one of the other words of the sentence functions as the non-verbal element of the compound verb. Thus, providing the machine with a ready-made list of compound verbs of the language substantially reduces the chance of misrecognition. Since the base structures of sentences rely on the valencies of verbs, one attempting to build a syntactic corpus of Persian needs to identify verbs in sentences first. Moreover, this resource will be valuable for the task of multi-word verb identification. This work is part of the Persian dependency corpus project. This led us to build the first Persian syntactic valency lexicon (the reasons why we chose dependency representation for Persian is stated in section 4). In addition, this lexicon can be enriched with semantic valencies as well as examples from the corpus that we are going to develop.

In this paper, we report the development of the first syntactic valency dictionary of Persian verbs. In section 2, some information about valency and DG is stated. Section 3 has a review of works in other languages. In section 4, our reasons for choosing dependency representation for Persian syntax is stated. Section 5 presents the steps toward creating the lexicon and some statistics related to it. Section 6 concludes the paper and proposes some new tasks to advance the work done here.

2. Valency and Dependency Grammar

Modern DG was first introduced in (Tesnière, 1953). In DG, it is assumed that words in a syntactic structure, have asymmetrical binary relations with each other (Kübler, McDonald, and Nivre, 2009). There are two main assumptions in this grammar. The first is that all sentences have a main verb and the second is that the obligatory and optional valencies of the verb determine the base structure of the sentence. The major difference between DG and generative grammar is the DG's reliance on words rather than phrases.

The most important notion in DG is known as valency. In DG, the main verb of the sentence bears the central gravity of the sentence and identifies the sentence base structure (Tesnière, 1980). The notion of valency is inspired from chemistry, in which each element has an ability to combine with atoms of other elements (Tesnière, 1980). Thus, verbs, nouns and adjectives have the ability to take certain dependents (some obligatory and some optional) the collection of which is known as valency (Tesnière, 1980).

3. Related Works

One of the earliest developments in building valency lexicons is the work done in (Grishman, Macleod, and Meyers, 1994). In (Baldwin, Bond, and Hutchinson, 1999), a bilingual valency dictionary is used for

[1] This paper is funded by Computer Research Center of Islamic Sciences (CRCIS).

Japanese-English machine translation. In (Herbst, Heath, and Roe, 2004), valencies of some verbs, nouns and adjectives in English enriched with examples and notes on their semantics, are collected. The frequency and hardship for foreign language learners were the criteria for selecting the words (Fillmore, 2009). In (Proisl and Kabashi, 2010), the mentioned resource is used for left-associative grammars and authors approve the value of its use for natural language processing tasks. Another valency dictionary for English is introduced in (Korhonen, Krymolowski, and Briscoe, 2006) where information about subcategorization is tagged. The other resource is FrameNet (Fillmore, Johnson, and Petruck, 2003), where some parts of British National Corpus (BNC) are annotated with phrase types, semantic and grammatical roles. The other project is VerbNet (Kipper, Dang, and Palmer, 2000), which is based on Levin's verb classes (Levin, 1993). In VerbNet, selectional preferences, verb senses and argument structure of each English verb is annotated. The other English valency dictionary is in (Semecký and Cinková, 2006), which is based on Functional Generative Grammar and extracted from Prague English dependency treebank.

There are also some other works on other languages, mostly Czech. In the Czech language, there are resources such as VerbaLex (Hlaváčková, et al., 2006) available which have syntactic annotation of verb valencies. In (Lopatková, Řezníčková, and Žabokrtský, 2006), verb valencies are extended to nouns. Considering semantic roles in FrameNet, (Kettnerová, Lopatková, and Hrstková, 2008), add semantic roles to Czech valency lexicon semi-automatically. The other Valency dictionary of Czech is PDT-Vallex (J. Hajič et al., 2003) extracted from Prague Dependency Treebank (PDT) (Bömová, Hajič, Hajičová, and Hladká, 2003). In addition, there are resources in other languages, such as French subcategorization lexicon (Messiant, Korhonen, and Poibeau, 2008), Romanian verb valency lexicon (Barbu, 2008), Arabic verb valency lexicon (Bielický and Smrz, 2008) extracted from Prague Arabic Dependency Treebank (PADT) (Jan Hajič, Smrž, Zemánek, Šnaidauf, and Beška, 2004), Croatian lexicon (Agic et al., 2010), German lexicon (Hinrichs and Telljohann, 2009) and a bank of Russian valencies (Lyashevskaya, 2010).

4. Persian Valency Lexicon[2]

One of the major requirements of Persian language processing, is the need for a syntactically annotated corpus. The Process of manually annotating a corpus needs a lot of time. For example, the first phase of Chinese treebank lasted five years (Hwa, Resnik, Weinberg, Cabezas, and Kolak, 2005). Hence, choosing a good representation for the corpus annotation is very important (Hwa, et al., 2005). There are many advantages to dependency representation to encourage one to select this type of representation. The first is its proximity to human interpretation (Kübler, et al., 2009). The second advantage is the appropriateness of this type of annotation for free-word order languages like Persian, i.e. with the non-projectivity in dependency trees free-word order can be represented much more easily. Witness to this claim is the choice of dependency annotation in Czech (Bömová, et al., 2003), Turkish (Oflazer, Say, Hakkani-Tür, and Tür, 2003), Danish (Kromann, 2003), German (Van der Beek, Bouma, Malouf, and Van Noord, 2002), Arabic (Jan Hajič, et al., 2004) and Latin (McGillivray, Passarotti, and Ruffolo, 2009). The third advantage is that dependency annotation can be automatically converted to phrase-based style, but the reverse is not completely possible (Johnson, 2007). These advantages led us to adopt a dependency representation.

One challenge toward collecting raw data for the treebank, is balancing the data for the purpose of data representativeness. Considering the importance of the contribution of the verb to the overall structure of the sentence, one needs to balance the data based on verb valency frequency distribution. The problem is that no valency dictionary for Persian verbs has ever been developed.

Furthermore, there is no comprehensive list of Persian compound verbs produced based on linguistic criteria. Although the task of manually identifying Persian compound verbs seems to be a simple task, the annotators faced many hardships in agreeing on whether a sequence of words is a real compound verb or not. These problems led the team to collect verb candidates from a large corpus and after omitting non-compound verbs to annotate their valencies. All of the above-mentioned steps took the team more than ten months to collect raw candidates of Persian verbs, omit non-verbs, annotate valencies, proofread the valencies and collect raw sentences based on verb valency distribution.

4.1. Verb Valency Types in Persian Language

Verb valency is the total number of complements a verb can take. This is an abstract notion and belongs to the mental lexicon of the native speaker of the language. Verb valency types demonstrate the possible forms in which the verb can be realized, however base structure is a more concrete notion which shows the actual realization of the verb complements in a sentence. In other words, any valency structure present in the lexicon can be realized as one or more base structures.

(Mohadjer-Ghomi, 1978) is the first work to address Persian verb valencies from the lexical valency theory. It specifies ten types of complements for verbs. These are as follows: 1) nominative object, 2) accusative object, 3) genitive object, 4) dative object, 5) prepositional object, 6) adverb of quantity, 7) adverb of direction, 8) number, 9) comparison, and 10) verbal complement.

Another work dealing with syntactic complements of verbs in Persian is (Ahadi, 2001). He identifies 11 syntactic complements for verbs. The complements are 1) subject, 2) direct complement, 3) pre-ezafe complement, 4) ezafe[3] complement, 5) ezafe complement followed by

[2] The valency lexicon is available at http://dadegan.ir/en/download

[3] The ezafe construction is an extremely productive means for modifying nouns as well as linking other nonverbal heads and their complements. The ezafe links a head noun to an adjective

the morpheme "ra", 6) enclitic complement, 7) place complement, 8) quantity complement, 9) nominal complement, 10) adjectival complement, and 11) verbal complement.

(Tabibzadeh, 2006) is an attempt to determine the syntactic complements of Persian verbs within DG. It enumerates eight kinds of syntactic complements for verbs. They are 1) subject, 2) object, 3) prepositional complement, 4) ezafe complement, 5) complement clause, 6) mosnad[4], 7) tamiz[5], and 8) adverbial complement. It also recognizes 23 base structures for Persian sentences. We add to the above mentioned list of complements "second object". We also add four base structures to those listed in (Tabibzadeh, 2006). Table 2 shows the base structures recognized by the authors of this paper. The last four ones differ from base structures in (Tabibzadeh, 2006). Table 1 gives the list of symbols and abbreviations used in table 2.

Abbreviation	Description
SBJ	Subject
NULL-SBJ	Null subject
VCL	Complement clause of verb
OBJ	Object
VPP	Prepositional complement of verb
EZC	Ezafe complement
MOS	Mosnad
AVDC	Adverbial clause
TAM	Tamiz
OBJ2	Second object
|| ||	Base structure

Table 1. Abbreviation used in Persian valency

Base Structures	
||SBJ||	||SBJ,VPP,VPP||
||NULL-SBJ,VCL||	||SBJ,VPP,EZC||
||SBJ,OBJ||	||SBJ,VPP,ADVC||
||SBJ,VPP||	||SBJ,VPP,VCL||
||SBJ,EZC||	||SBJ,VPP,TAM||
||SBJ,VCL||	||SBJ,EZC,VCL||
||SBJ,MOS||	||NULL-SBJ,VCL,VPP||
||SBJ,ADVC||	||SBJ,OBJ,VPP,VPP||
||SBJ,OBJ,VPP||	||SBJ,VPP,VPP,ADVC||
||SBJ,OBJ,EZC||	||SBJ,ADVC,EZC||
||SBJ,OBJ,VCL||	||SBJ,OBJ,VPP,TAM||
||SBJ,OBJ,TAM||	||SBJ,OBJ,VPP,ADVC||
||SBJ,OBJ,MOS||	||SBJ,OBJ,VPP,OBJ2||
||SBJ,OBJ,ADVC||	||SBJ,VPP,VPP||

Table 2. Possible Base Structures in Persian

5. Steps for Creating the Lexicon

Several preprocessing steps were taken for building the lexicon. In the first step, a POS tagger was built from the data in Bijankhan corpus (Bijankhan, 2004). In the second step, all texts in Bijankhan corpus and Hamshahri corpus (Aleahmad, Amiri, Rahgozar, and Orumchian, 2009) (a collection of raw texts with subject category for each document), were tagged and lemmatized. In the third step, after lemmatizing verbs and converting them into their lemmas, association measures for Persian compound verb identification in (Rasooli, Faili, and Minaei-Bidgoli, 2011) were used to find probable candidates of Persian compound verbs. The baseline association measure in this work is pointwise mutual information (PMI) measure. After finishing all of the steps mentioned, the verb candidates were manually tagged with valency types or omitted (because of not being a compound verb). The final results were proofread four times by several annotators, in order to be sure of the output results. In Fig. 1, a sample output of the lexicon in the XML format is shown. The statistics of the lexicon is shown in Table 3.

```
<Verb-plus-Valency>
<Verb>
<Past-light-verb>کرد</Past-light-verb>
<Present-light-verb>کن</Present-light-verb>
<Prefix>-</Prefix>
<Nonverbal-element>صحبت</Nonverbal-element>
<Preposition>-</Preposition>
</Verb>
<Valency>SBJ,(VPP)[با],(VPP)[در خصوص|در مورد|درباره|از]
</Valency>
<Valency>SBJ,EZC[RA+/-]</Valency>
</Verb-plus-Valency>
```

Fig. 1: A sample output of the valency lexicon

6. Conclusion and Further Works

There are many open tasks that can enrich this lexicon such as adding semantic frames, adding sample sentences from corpus to the lexicon and extending the project to noun and adjective valencies. Using datasets such as VerbNet and a bilingual dictionary of Persian-English verbs will add semantic frames to this lexicon automatically, reducing the expenses.

Number of distinct verbs	4282
Number of valencies	5429
Average distinct valency per verb	1.268
Maximum number of valency per verb	5

Table 3. Statistics in the valency lexicon

Acknowledgements

This paper is funded by Computer Research Center of Islamic Sciences (CRCIS). We also like to thank Neda Poormorteza-Khameneh, Akram Shafie, Saeedeh Ghadrdoost-Nakhchi, and Farzaneh Bakhtiary for their

(phrase), noun (phrase), adverb (phrase), prepositional phrase or infinitive. The ezafe can also link adjective, quantifier and prepositional heads to their complements (Mahootian and Gebhardt, 1997).

[4] Mosnad is a property of a noun, an adjective or a pronoun ascribed to the subject of a sentence whose main verb is a predicative verb such as verb forms derived from any of these Persian infinitives: "budæn" (= to be), "šodæn" (=to become), "ʔæstæn" (= to be), etc.

[5] Tamiz is a property of an adjective or a noun ascribed to the object by the subject of a sentence whose main verb is one such as "namidæn" (= to name), "xandæn" (= to call), "danestæn" (= to consider), etc.

help on annotation. We also want to thank Dr. Omid Tabibzadeh and his student, Sara Kalali, and members of the Scientific Society for the Students of Linguistics, University of Tehran. We also appreciate Dr. Sandra Kübler and Professor Joakim Nivre for their helpful answers to our questions.

References

Agic, x, Z., x030C, K. ojat, Tadic, and M. (2010). An experiment in verb valency frame extraction from Croatian Dependency Treebank. Paper read at Information Technology Interfaces (ITI), (2010) 32nd International Conference on, 21-24 June 2010.

Ahadi, Shahram. (2001). verbergänzungen und zusammengesetzte in Persischen, Eine Valenzthoretische Analyse: Dr. Ludvig Reichert, Verlag, Wiesbaden.

Aleahmad, Abolfazl, Hadi Amiri, Masoud Rahgozar, and Farhad Orumchian. (2009). Hamshahri: A Standard Persian Text Collection. Knowledge-Based Systems 22 (5):382-387.

Baldwin, Timothy, Francis Bond, and Ben Hutchinson. (1999). A valency dictionary architecture for machine translation. In Eighth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99. Chester, UK.

Barbu, Ana-Maria. (2008). First Steps in Building a Verb Valency Lexicon for Romanian. In Text, Speech and Dialogue, edited by P. Sojka, A. Horák, I. Kopecek and K. Pala: Springer Berlin / Heidelberg.

Bielický, V., and O. Smrz. (2008). Building the Valency Lexicon of Arabic Verbs. In 6th Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco.

Bijankhan, Mahmood. (2004). The role of the corpus in writing a grammar: An introduction to a software. Iranian Journal of Linguistics 19 (2).

Bömová, Alena, Jan Hajič, Eva Hajičová, and Barbora Hladká. (2003). The Prague Dependency Treebank: A three-level annotation scenario. Treebanks: Building and Using Parsed Corpora:103-127.

Fillmore, C.J. (2009). A Valency Dictionary of English. International Journal of Lexicography 22 (1):55.

Fillmore, C.J., C.R. Johnson, and M.R.L. Petruck. (2003). Background to FrameNet. International Journal of Lexicography 16 (3):235.

Grishman, Ralph, Catherine Macleod, and Adam Meyers. (1994). Comlex Syntax: building a computational lexicon. In 15th conference on Computational linguistics - Volume 1. Kyoto, Japan: Association for Computational Linguistics.

Hajič, J., J. Panevová, Z. Urešová, A. Bémová, V. Kolárová, and P. Pajas. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. Paper read at The Second Workshop on Treebanks and Linguistic Theories

Hajič, Jan, Otakar Smrž, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. (2004). Prague Arabic Dependency Treebank: Development in data and tools. In Proceedings of the NEMLAR 2004 International Conference on Arabic Language Resources and Tools. Cairo, Egypt.

Herbst, T., David Heath, and Ian F. Roe. (2004). A Valency Dictionary of English: A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns, and Adjectives. Edited by D. Götz. Vol. 40: Walter de Gruyter.

Hinrichs, E., and H. Telljohann. (2009). Constructing a Valence Lexicon for a Treebank of German. In Seventh International Workshop on Treebanks and Linguistic Theories. Utrecht, Netherlands.

Hlaváčková, Dana, Aleš Horák, and Vladimír Kadlec. (2006). Exploitation of the VerbaLex Verb Valency Lexicon in the Syntactic Analysis of Czech. In Text, Speech and Dialogue, edited by P. Sojka, I. Kopecek and K. Pala: Springer Berlin / Heidelberg.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. (2005). Bootstrapping parsers via syntactic projection across parallel texts. Natural Language Engineering 11 (3):311-325.

Johnson, Mark. (2007). Transforming projective bilexical dependency grammars into efficiently parseable CFGs with unfold-fold. In Proceeding of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association of Computational Linguistics.

Kettnerová, Václava, Markéta Lopatková, and Klára Hrstková. (2008). Semantic Roles in Valency Lexicon of Czech Verbs: Verbs of Communication and Exchange. In Advances in Natural Language Processing, edited by B. Nordström and A. Ranta: Springer Berlin / Heidelberg.

Kipper, K., H.T. Dang, and M. Palmer. (2000). Class-based construction of a verb lexicon. In Seventeenth National Conference on Artificial Intelligence (AAAII-2000). Austin, Texas.

Korhonen, Anna, Yuval Krymolowski, and Ted Briscoe. (2006). A large subcategorization lexicon for natural language processing applications. In 5th International Conference on Language Resources and Evaluation (LREC'05).

Kromann, M.T. (2003). The Danish Dependency Treebank and the DTAG treebank tool. In Second Workshop on Treebanks and Linguistic Theories (TLT 2003). Växjö, Sweden.

Kübler, Sandra, Ryan McDonald, and Joakim Nivre. (2009). Dependency Parsing. Edited by G. Hirst, SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES: Morgan and Claypool.

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

Lopatková, Markéta, Veronika Řezníčková, and Zdeněk Žabokrtský. (2006). Valency Lexicon for Czech: From Verbs to Nouns. In Text, Speech and Dialogue, edited by P. Sojka, I. Kopecek and K. Pala: Springer Berlin / Heidelberg.

Lyashevskaya, O. (2010). Bank of Russian Constructions and Valencies In Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta.

Mahootian, M., and Gebhardt, L., (1997). Persian Descriptive Grammars, Taylor and Francis Routledge.

McGillivray, B., M. Passarotti, and P. Ruffolo. (2009). The Index Thomisticus treebank project: Annotation, parsing and valency lexicon. Traitement Automatique des Langues 50 (2).

Messiant, Cédric, Anna Korhonen, and Thierry Poibeau. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs In Language Resources and Evaluation Conference (LREC). Marrakech, Morocco.

Mohadjer-Ghomi, Siamak. (1978). Eine kontrastive Untersuchung der Satzbaupläne im Deutschen und Persischen: Burg-Verlag.

Oflazer, K., B. Say, D.Z. Hakkani-Tür, and G. Tür. (2003). Building a Turkish treebank. Treebanks: Building and Using Parsed Corpora 20:261-277.

Proisl, T., and B. Kabashi. (2010). Using High-Quality Resources in NLP: The Valency Dictionary of English as a Resource for Left-Associative Grammars. In Seventh conference on International Language Resources and Evaluation (LREC'10), edited by N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner and D. Tapias. Valletta, Malta: European Language Resources Association (ELRA).

Rasooli, Mohammad Sadegh, Heshaam Faili, and Behrouz Minaei-Bidgoli. (2011). Unsupervised Identification of Persian Compound Verbs. In Submitted in 10th Mexican International Conference on Artificial Intelligence (MICAI).

Semecký, Jiří, and Silvie Cinková. (2006). Constructing an English valency lexicon. In Workshop on Frontiers in Linguistically Annotated Corpora 2006. Sydney, Australia: Association for Computational Linguistics.

Tabibzadeh, Omid. (2006). Verb Valency and Sentence Base Structures in Modern Persian. Tehran, Iran: Markaz.

Tesnière, Lucien. (1953). Esquisse d'une Syntaxe structurale. Paris: Klincksieck.

Tesnière, Lucien. (1980). Grundgzüge der Strukturalen Syntax. Edited by H. V. U. Engel. Stuttgart: Klettcotta.

Van der Beek, L., G. Bouma, R. Malouf, and G. Van Noord. (2002). The Alpino dependency treebank. Language and Computers 45 (1):8-22.