Extending the coverage of a MWE database for Persian CPs exploiting valency alternations

Pollet Samvelian[1], Pegah Faghiri[1], Sarra El Ayari[2]
[1] Université Sorbonne Nouvelle & UMR Mondes iranien et indien (CNRS)
[2] Université Paris Diderot & Labex EFL
E-mail: pollet.samvelian@univ-paris3.fr, pegah.faghiri@univ-paris3.fr, sarra.elayari@univ-paris-diderot.fr

Abstract

PersPred is a manually elaborated multilingual syntactic and semantic Lexicon for Persian Complex Predicates (CPs), referred to also as "Light Verb Constructions" (LVCs) or "Compound Verbs". CPs constitutes the regular and the most common way of expressing verbal concepts in Persian, which has only around 200 simplex verbs. CPs can be defined as multi-word sequences formed by a verb and a non-verbal element and functioning in many respects as a simplex verb. Bonami & Samvelain (2010) and Samvelian & Faghiri (to appear) extendedly argue that Persian CPs are MWEs and consequently must be listed. The first delivery of PersPred, contains more than 600 combinations of the verb *zadan* 'hit' with a noun, presented in a spreadsheet. In this paper we present a semi-automatic method used to extend the coverage of PersPred 1.0, which relies on the syntactic information on valency alternations already encoded in the database. Given the importance of CPs in the verbal lexicon of Persian and the fact that lexical resources cruelly lack for Persian, this method can be further used to achieve our goal of making PersPred an appropriate resource for NLP applications.

Keywords: Complex Predicates, Multiword Expressions, Persian

1. Introduction

In this paper we present a semi-automatic method used to extend the coverage of a syntactic and semantic database for Persian CPs, namely PersPred, whose first delivery contained around 600 manually encoded entries. The method relies on the syntactic information concerning valency alternations already present in the database. It gave accurate results for more than 80% of entries, allowing thus to extend the coverage of PersPred by 70%. Given the importance of CPs in the verbal lexicon of Persian and the fact that lexical resources cruelly lack for Persian, this method can be further used to achieve our goal of making PersPred an appropriate resource for NLP applications.

2. PersPred : A database for Persian CPs

PersPred is the first manually annotated syntactic and semantic database for Persian complex predicates (CPs). The latter constitute the regular and the most common way of expressing verbal concepts in Persian, which has only around 200 simplex verbs. CPs can be defined as a multi-word sequence formed by a verb and a non-verbal element and functioning in many respects as a simplex verb (or predicate), e.g. harf zadan 'to talk' (Lit. 'talk hit'), bâz kardan 'to open' (Lit. 'open do'), bar dâštan 'to take' (Lit. 'PARTICLE have'), be kâr bordan 'to use' (Lit. 'to work take'). These sequences are also referred to as "Light Verb Constructions" (Karimi-Doostan 1997).

Bonami & Samvelain (2010), Samvelian (2012) and Samvelian & Faghiri (to appear) extendedly argue that Persian CPs are MWEs and as such need to be listed. However, despite several attempts, this task has not been carried out in a systematic way and a large-scale lexical resource for Persian CPs is cruelly missing (Taslimipoor et al. 2012).

PersPred, the first syntactic and semantic database of Persian CPs, aims to contribute to fill this gap. Its first delivery, PersPred 1, contains 648 CPs formed by the combination of the verb zadan 'to hit' and a nominal element (Samvelian & Faghiri, 2013). The main interest of PersPred is its rich semantic and syntactic annotation. Indeed, for each element 22 fields relating to different lexical, syntactic and semantic information are annotated: 9 fields provide information on the lemma of the CP and its combining parts, including French and English translations of the Noun, the Verb and the CP, 5 fields are dedicated to semantic information, e.g. the semantic class and the type of meaning extension (metaphor, metonymy, synecdoche) if applicable, and finally 8 fields represent the syntactic construction of the CP and its English equivalent through an abstract syntactic template inspired by Gross (1977). Valency alternations and synonymy are represented through 3 fields, Intransitive, Transitive and Synonymous Variants. Table 1 below illustrates these fields via the example of the CP *âb zadan* 'to wet'. Note that 2 extra fields provide (at least) one attested example in Persian script and its phonetic transcription. (cf. Samvelian & Faghiri (2013) for a detailed presentation of PersPred).

One of the main difficulties when developing PersPred1 was the time-consuming task of annotation, especially regarding the semantic and syntactic information. Note that the number of lexicalized Persian CPs probably exceeds a few thousand. Rassooli et al. (2011), for instance, provide a list of 4000 "verbs", including simplex

verbs and CPs, which is far from being exhaustive. Elaborating methods for a semi-automatic extension of the coverage of PersPred has thus been one of our priorities since the delivery of PersPred 1.

One of the methods we have explored in order to extend the coverage of the database with the least human intervention is a semi-automatic method exploiting the already annotated information. More precisely, on the basis of the existing entries and the information on valency alternations and synonymy, we have created and pre-annotated new entities. This pre-annotations need only to be validated by an annotator in order to be integrated into the database. Before presenting this bootstrapping method we briefly introduce the syntactic annotation provided by PersPred.

Lemma information	
Verb	آب
Noun	زدن
N-transcription	âb
V-transcription	zadan
CP-lemma	âb-zadan0
N-FR-translation	eau
N-EN-translation	water
CP-FR-translation	mouiller
CP-EN-translation	to wet
Subcategorization and syntactic information	
Synt-Construction	N0 Prep N1 N2 V
PRED-N	N2
Prep-N1	be
Prep-N2	NONE
Construction-trans-En	N0 wets N2
Intrans-Var	xordan
Trans-Var	NONE
Syn-Var	NONE
Semantic information	
Sem-Class	Spreading
Sem-Super-Class	Locatum
Constant-Sem	Liquid
Subject-Sem	Human
Meaning-Exension	NONE

Table 1 : Different fields of PersPred illustrated for *âb zadan*

3. The syntactic annotation in PersPred

In PersPred, the subcategorization frame is provided through an annotation inspired by Gross (1977). To give an example, for the CP *ab zadan* 'to water' (Lit. 'water hit'), the syntactic structure can be N0 Prep N1 N2 V or N0 N1 N2 V; where N stands for a bare noun or a nominal projection (i.e. NP) and the number following N indicates the obliqueness hierarchy among nominal elements: N0 is the first argument (subject); N1 the direct object; Prep N1 the prepositional object and so on.

In addition to the subcategorization frame, PersPred encodes information on valency alternations and synonymy for each CP. The value of these features is either a verbal lemma or NONE, if there is no attested variant. Intrans-Variant provides the lemma of one or several verbs that can be used to produce a CP where the Patient (N1 or N2) argument is assigned the subject function, i.e. becomes N0. This alternation is somehow comparable to the passive alternation. Trans-Variant gives the lemma of the verb(s) used to add an extra argument (or participant) to the CP. This external participant generally has a Cause interpretation and is realized as the subject of the "transitive/Causative" CP. The first argument of the initial CP is mapped in this case onto the Object function. Syn-Variant gives the lemma of the set of verbs forming a synonymous predicate with the same noun.

4. Our bootstrapping method for extending the coverage of PersPred

Our method consists in using the information encoded by the set of Variant features to generate new entries. In this paper, we illustrate this method for *zadan*-CPs. The same method can be used to extend the coverage of PersPred with other verbs, e.g. *gereftab* 'to take' and *dâdan* 'to give', once they are integrated into the database.

4.1 Intransitive variants

260 entries have at least one intransitive variant. The intransitive variant can correspond to one or several lemmas among the following: *xordan* 'to collide', *didan* 'to see', *gereftan* 'to take' or *residan* 'to arrive'. The most frequent intransitive variant is *xordan* which is available for 246 entries. The second frequent variant is *didan* with 12 entries. We did not include the two other variants because they involved only a very small number of entries.

Note that, given the possibility to have more than one variant available for a CP, some CPs can potentially give rise to more than one pre-annotated entry. It should be noted that this possibility is not limited to the intransitive alternation and in the case of other variants (i.e. transitive and synonymous) as well, a number of CPs can give rise to more than one pre-annotated entries.

Z*adan*-CPs that allow the *xordan*-alternation have the two following syntactic constructions (cf. Synt-Construction in table 1):

(1) N0 N1 N2 V
(2) N0 Prep N1 N2 V

N.B. In both constructions N2 corresponds to the nominal element of the CP.

For (1), the link between the transitive and intransitive alternation is straightforward:

(3) *Zadan*-CP (trans): N0 (Agent) N1 (Patient) N2 V
Xordan-CP (intrans): ~~N0 (Agent)~~ N1 (Patient) N2 V
N0 (Patient) N1 V

The argument corresponding to the subject in the *zadan*-CP, namely N0, is suppressed in the *xordan*-CP and the argument corresponding to the DO, the Patient, namely N1, is mapped into the subject. In the *xordan*-CP, N1 corresponds to the nominal element of the CP. The following pairs illustrate this situation:

(4a) *kotak zadan* → *kotak xordan*
'to beat' → 'to be beaten'
N0 N1 N2 V → N0 N1 V

(4b) *rang zadan* → *rang xordan*
'to paint' → 'to be painted'
N0 N1 N2 V → N0 N1 V

For (2), the link between the two constructions is slightly more complex, because two different realizations for the intransitive construction are possible. Besides the construction similar to the one in (3) where the patient, i.e. the argument corresponding to the prepositional complement, is mapped into the subject, there is also the possibility to realize this argument as the prepositional complement and the nominal element of the CP, namely N2, as the subject of *xordan*-CP. These two possibilities are given in (5a) and (5b) and illustrated by the two pairs in (6) and (7) below.

(5a) *Zadan*-CP: N0 (Agent) Prep N1 (Patient) N2 V
Xordan-CP: ~~N0 (Agent)~~ N1 (Patient) N2 V
→ **N0 (Patient) N1 V**

(5b) *Zadan*-CP: N0 (Agent) Prep N1 (Patient) N2 V
Xordan-CP: ~~N0 (Agent)~~ Prep N1 (Patient) N2 V
Prep N1 (Patient) N0 V
→ **N0 Prep N1 (Patient) V**

(6) *âsib zadan* → *âsib xordan*
'to damage' → 'to be damaged'
a. N0 Prep N1 N2 V → N0 N1V
b. N0 Prep N1 N2 V → N0 Prep N1 V

(7) *sadame zadan* → *sadame xordan*
'to harm' → 'to be harmed'
a. N0 Prep N1 N2 V → N0 N1 V
b. N0 Prep N1 N2 V → N0 Prep N1 V

For the lexical fields, namely the French and English translations of the CP, we generated the respective corresponding passive forms with the rule in (8) as illustrated by the pair in (9).

(8) *Zadan*-CP → *Xordan*-CP
(EN) $V_{INFINIVE}$ → 'be + $V_{PAST\text{-}PARTICIPLE}$'
(FR) $V_{INFINIVE}$ → 'être + $V_{PAST\text{-}PARTICIPLE}$'

(9) *Zadan*-CP → *Xordan*-CP
sadame zadan → *sadame xordan*
(EN) 'to harm' → 'to be harmed'
(FR) 'endommager' → 'être endommagé'

And finally, for the English mapping of the syntactic construction (cf. Construction-trans-En in table 1) we used the rule in (10). N.B. The English mapping is identical for the different possible structures of each CP.

(10) *Zadan*-CP → *Xordan*-CP
N0 $V_{PRES.3.SG}$ N1 → N0 is $V_{PAST\text{-}PARTICIPLE}$

Examples (11) and (12) illustrate the result of these transformations for *kotak zadan* and *sadame zadan*.

(11) *kotak zadan* → *kotak xordan*
'to beat' → 'to be beaten'
'frapper' → 'être frappé'
N0 N1 N2 V → N0 N1 V
'N0 beats N1' → 'N0 is beaten'

(12) *sadame zadan* → *sadame xordan*
'to harm' → 'to be harmed'
'endommager' → 'être endommagé'
1. N0 N1 N2 V → N0 N1 V
'N0 harms N1' → 'N0 is harmed'
2b. N0 Prep N1 V → N0 Prep N1 V
'N0 harms N1' → 'N1 is harmed'

Note that for some *zadan*-CPs, like *sadame zadan* in (12) above, both constructions (1) and (2) are available. In this cases, the mapping rule in (3), applied to the construction (1), and the one in (5a), applied to the construction (2), produce the same results.

For 12 entries the intransitive variant can be formed by the verb *didan* 'to see'. It should be noted that these entries have both *xordan* and *didan* as their intransitive variant. The mapping rules for *didan*-alternation are the same as *xordan*-alternation, except that *didan*-CPs have only one possible realization, i.e. the prepositional realization (cf. 5b above) available for *xordan*-CPs is not possible in the case of *didan*. Example (14) illustrates the results of *didan*-alternation for *âsib zadan* and *sadame zadan.*

(14a) *âsib zadan* → *âsib didan*
'to damage' → 'to be damaged'
'endommager' → 'être endommagé'
N0 Prep N1 N2 V → N0 N1 V
'N0 damages N1' → 'N0 is damaged'

(14b) *sadame zadan* → *sadame didan*
'to harm' → 'to be harmed'
'endommager' → 'être endommagé'
N0 Prep N1 N2 V → N0 N1 V
'N0 harms N1' → 'N0 is harmed'

The remaining fields, except for the examples, are copied into the new entry.

4.2 Synonymous variants

267 entries have at least one synonymous variant; the most common synonymous verb is *kardan.* Contrary to the previous case, the verbs used to form synonymous variants display an important variety and no one is singled out as the regular alternate variant for *zadan*. The most frequent synonymous alternates are *kardan* 'to do', 79 entries, and, *kešidan* 'to pull', 63 entries. Other available variants, e.g. *resândan* 'to convey', *oftâdan* 'to fell', *andâxtan* 'to throw', concern only a very small number of entries and we did not include them in the operation.

For synonymous variants we decided to resort to a rather elementary rule and generated a new entry for each existing entry and each synonymous variant available. The rule to generate new entries consists of replacing *zadan* by the given variant. In other words, the rule only changes the fields related to the verb and copis the other fields, except for the examples.

4.3 Transitive variants

There are only 14 entries which have a transitive variant. This is not surprising, since most of *zadan*-CPs are transitive themselves. Given the limited number of the set, we decided not to resort to semi-automatic annotation in this case.

5. Manual validation

The semi-automatic method produced 402 pre-annotated entries, i.e. 260 and 142 for the intransitive and the synonymous operation respectively. We then manually validated each of these entries. In the case of the intransitive variants, 86% of the new entries were acceptable and only some small modifications were needed, namely, for French and English translation and the English mapping. As for the synonymous variants, almost all new entries were acceptable. However, some amount of modification was necessary, particularly for the fields related to the syntactic construction.

6. Conclusion

To conclude, we want to emphasize the need for manually elaborated resources with rich semantic and syntactic information. Even though time consuming at the first stage, they are extremely useful in the further development and extension of the initial resource.

7. Acknowledgements

This work was supported by the bilateral project PerGram, funded by the French National Research Agency (ANR) and the DGfS (Germany) [grant no. MU 2822/3-I] and a public grant funded by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

8. Reference

Bonami, Olivier and Samvelian, Pollet, 2010. Persian complex predicates: Lexeme formation by itself. Paper presented at Septièmes Decembrettes Morphology Conference, Toulouse

Gross, Maurice, "On the relations between syntax and semantics", Formal Semantics of Natural Language, E.L. Keenan (ed.), Cambridge: Cambridge University Press, 1975, pp. 389-405.

Karimi-Doostan, G. 1997. Light verb constructions in Persian. Essex University, England. Ph.D.Diss.

Pollet Samvelian (2012). Grammaire des prédicats complexes. Les constructions nom-verbe, Paris: Hermès-Lavoisier.

Pollet Samvelian and Pegah Faghiri (2013). Introducing PersPred, a syntactic and semantic database for Persian complex predicates. In Proceedings of the 9th Workshop on Multiword Expressions, Atlanta, Georgia, USA. Association for Computational Linguistics, pages 11-20.

Pollet Samvelian and Pegah Faghiri (to appear), Rethinking Compositionality in Persian Complex Predicates. in Proceedings of Berkeley Linguistics Society 39th Annual Meeting, February 16-17, 2013, Berkeley.

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, & Behrouz Minaei Bidgoli. (2011). A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. in 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics (pp. 227–231). Poznań, Poland.

Shiva Taslimipoor, Afsaneh Fazly, and Ali Hamzeh.2012. Using noun similarity to adapt an acceptability measure for persian light verb constructions. In Language Resources and Evaluation Conference (LREC 2012), Istanbul.