Introducing *PersPred*, a Syntactic and Semantic Database for Persian Complex Predicates

Pollet Samvelian and **Pegah Faghiri**
Université Sorbonne Nouvelle & CNRS
18, rue des Bernardins
75005, Paris, France
{pollet.samvelian, pegah.faghiri}@univ-paris3.fr

Abstract

This paper introduces *PersPred*, the first manually elaborated syntactic and semantic database for Persian Complex Predicates (CPs). Beside their theoretical interest, Persian CPs constitute an important challenge in Persian lexicography and for NLP. The first delivery, *PersPred 1*[1], contains 700 CPs, for which 22 fields of lexical, syntactic and semantic information are encoded. The semantic classification *PersPred* provides allows to account for the productivity of these combinations in a way which does justice to their compositionality without overlooking their idiomaticity.

1 Introduction

Persian has only around 250 simplex verbs, half of which are currently used by the speech community[2]. The morphological lexeme formation process outputting verbs from nouns (e.g. *xâb* 'sleep' > *xâb-idan* 'to sleep'; *raqs* 'dance' > *raqs-idan* 'to dance'), though available, is not productive. The verbal lexicon is mainly formed by syntactic combinations, including a verb and a non-verbal element, which can be a noun, e.g. *harf zadan* 'to talk' (Lit. 'talk hit'), an adjective, e.g. *bâz kardan* 'to open' (Lit. 'open do'), a particle, e.g. *bar dâštan* 'to take' (Lit. 'PARTICLE have'), or a prepositional phrase, e.g. *be kâr bordan* 'to use' (Lit. 'to work take'). These combinations are generally referred to as Complex Predicates (CPs), Compound Verbs or Light Verb Constructions (LVCs).

New "verbal concepts" are regularly coined as complex predicates (CPs) rather than simplex verbs, for instance *yonize kardan* 'to ionize' (Lit. 'ionized do') instead of *yon-idan*[3].

Several studies have focused on the dual nature of Persian CPs, which exhibit both lexical and phrasal properties (Goldberg, 2003; Vahedi-Langrudi, 1996; Karimi, 1997; Karimi-Doostan, 1997; Megerdoomian, 2002, among others). Indeed, these combinations display all properties of syntactic combinations, including some degree of semantic compositionality, which makes it impossible to establish a clearcut distinction between them and "ordinary" verb-object combinations for instance (cf. 2.1). On the other hand, these sequences also have word-like properties, since CP formation has all the hallmarks of a lexeme formation process, such as lexicalization (cf. 2.2). Thus, in the same way as the verbal lexicon of English includes all its simplex verbs, the inventory of the verbal lexicon in Persian, and consequently dictionaries, must include these com-

[1] *PersPred 1* is freely available under the LGPL-LR license, http://www.iran-inde.cnrs.fr/ (Language Resources for Persian).

[2] Sadeghi (1993) gives the estimation of 252 verbs, 115 of which are commonly used. Khanlari (1986) provides a list of 279 simplex verbs. The Bijankhan corpus contains 228 lemmas.

[3] In reality, there are verbs formed from nouns or adjectives, but they are mainly created by the Academy of Persian Language and Literature, who suggests and approves equivalents for the foreign general or technical terms. The verb *râyidan* 'to compute', for instance, is a recent creation by the Academy. However, it should be noted that these creations, which are far less numerous than spontaneous creations, are not easily adopted by native speakers, who almost systematically prefer using the CP counterpart, e.g. *kampyut kardan* (Lit. 'computation do') instead of *râyidan*.

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 11–20,
Atlanta, Georgia, 13-14 June 2013.

binations. However, despite several attempts, this task has not been carried out in a systematic way and such a resource is cruelly missing. Although dictionaries mention some of the lexicalized combinations, either under the entry associated to the verb, or to the non verbal element, the underlying criteria in the choice of combinations is far from being clear and the resulting list significantly varies from one dictionary to another.

Computational studies have also mentioned the lack of large-scale lexical resources for Persian and have developed probabilistic measures to determine the acceptability of the combination of a verb and a noun as a CP (Taslimipoor et al., 2012).

PersPred is a syntactic and semantic database, which aims to contribute to fill this gap by proposing a framework for the storage and the description of Persian CPs. Its first delivery, *PersPred 1.*, contains more than 700 combinations of the verb *zadan* 'hit' with a noun, presented in a spreadsheet.

PersPred is not only a lexicographic resource, it is also the implementation of a theoretical view on Persian CPs. Adopting a Construction-based approach (cf. 4), *PersPred* sheds a new light on some crucial and closely related issues in CP formation:

- The way the productivity of these combinations can be accounted for despite their idiomaticity and the link generally established between compositionality and productivity (cf. 3).

- The relation between "lexical" and "light" verbs and the validity of such a distinction for a great number of Persian verbs.

The fact that Persian has *only* around 250 simplex verbs has a very obvious consequence which has generally been overlooked by theoretical studies: Almost *all* Persian verbs are light verbs, or, more precisely, are simultaneously light and lexical verbs. In other words, if one establishes a scale of specificity in the verbal meaning (Ritter and Rosen, 1996) going from highly specific verbs (e.g. *google*, *milk*) to lowly specific ones (e.g. *do*, *make*), most Persian verbs are located somewhere in the middle of the scale. Consequently, in many CPs, the verb has a lexical semantic content and cannot be considered as a light verb *sensu stricto*. This also entails that Persian CPs are not always as idiomatic as English LVCs, for instance, and that many aspects of their formation can be accounted for via compositionality. By providing a fine-grained semantic classification for Persian CPs, *PersPred* proposes a solution that does justice to the compositionality of these combinations, thus allowing to account for their productivity.

2 Persian CPs as Multiword Expressions

Several studies, including those in computational linguistics, treat Persian CPs like LVCs in languages such as English and French, and thus as MWEs (Fazly et al., 2007, among others). However, the fact that Persian CPs are generally formed by a "bare" (non-determined, non-referential) noun and a verb, in an adjacent position, makes them far more cohesive than English LVCs for instance, and leads some studies to treat these combination as *words* by default (Goldberg, 1996).

2.1 Phrasal Properties

It has been shown by several studies (Karimi-Doostan, 1997; Megerdoomian, 2002; Samvelian, 2012) that the two elements in a CP are clearly separate syntactic units: a) All inflection is prefixed or suffixed on the verb, as in (1), and never on the noun. b) The two elements can be separated by the pronominal clitics, (2), the future auxiliary, (3), or even by clearly syntactic constituents, (4). c) Both the noun and the verb can be coordinated, (5) and (6) respectively. d) The noun can be extracted, (7). e) CPs can be passivized, (8). In this case, the nominal element of the CP can become the subject of the passive construction, as does the Direct Object of a transitive construction. f) Finally, the noun can head a complex NP, (9).

(1) Maryam bâ Omid harf **ne-mi**-zan-**ad**
Maryam with Omid talk NEG-IPFV-hit-3S
'Maryam does not talk to Omid.'[4]

(2) Dust=**aš** dâr-am
friend=3S have-1S
'I like her/him/it.'

[4] DDO = definite direct object marker; EZ = *Ezafe* particle; IPFV = imperfective, NEG = negation, PP = past participle.

(3) Maryam Omid=râ dust **xâh-ad** dâšt
Maryam Omid=DDO friend AUX-3S had
'Maryam will like Omid.'

(4) Dast **be begol-hâ** na-zan
hand to flower-PL NEG-hit
'Don't touch the flowers.'

(5) Mu-hâ=yaš=râ **boros** yâ **šâne** zad
hair-PL=3S=DDO brush or comb hit
'(S)he brushed or combed her hair.'

(6) Omid sili **zad** va **xord**
Omid slap hit and strike
'Omid gave and received slaps.'

(7) **Dast** goft-am be gol-hâ ___ na-zan
hand said-1S to flower-PL ___ NEG-hit
'I told you not to touch the flowers.'

(8) a. Maryam be Omid tohmat zad
Maryam to Omid slander hit
'Maryam slandered Omid.'

b. Be Omid tohmat zade šod
to Omid slander hit.PP become
'Omid was slandered.'

(9) [In **xabar**=e mohem]=râ be mâ **dâd**
this news=EZ important=DDO to us gave
'(S)he gave us this important news.'

These observations show that the syntactic properties of CPs are comparable to regular Object-Verb combinations. While the noun in a CP is more cohesive with the verb than a bare direct object (in terms of word order, differential object marking, pronominal affix placement), it is impossible to draw a categorical syntactic distinction between the two types of combinations.

2.2 Lexical and Idiomatic Properties

While clearly being syntactic combinations, Persian CPs display several lexeme like properties (Bonami and Samvelian, 2010). From a semantic point of view, their meaning can be unpredictable (i.e. conventional). From a morphological point of view, the whole sequence behaves like a word in the sense that it feeds lexical formation rules. Finally, the association of a given noun and a given verb is more or less idiomatic.

CPs are lexicalized. In many cases, the meaning of a CP is not fully predictable from the meaning of its components. N-V combinations are subject to various levels of lexicalization.

In some cases, the CP meaning is a **specialization** of the predictable meaning of the combination. For instance *čâqu zadan* 'to stab' (Lit. 'knife hit') is not only to hit somebody with a knife; *dast dâdan* 'to shake hands' (Lit. 'hand give') does not only imply that you give your hand to somebody; *âb dâdan*, 'to water' (Lit. 'water give') is not just pouring water on something; šir dâdan 'to breastfeed' (Lit. 'milk give') is not just the action of giving milk to somebody. These particular specializations have to be learned, in the same way as one has to learn the meaning of the verbs such as *water* or *towel* in English.

In other examples **semantic drift** has taken place, either by metaphor or by metonymy. The link between the compositional meaning and the lexicalized meaning is sometimes still recoverable synchronically. For instance, the lexicalized meaning of *guš kardan* 'to listen' (Lit. 'ear do') can be recovered via metonymy. The CP designates the prototypical action done by ears. Likewise, in *zanjir zadan* 'to flagellate' (Lit. 'chain hit'), the elliptical element of the meaning, *pošt* 'shoulder', can also be recovered. The CP comes in fact from *bâ zanjir (be) pošt zadan* 'to hit one's shoulders with chains'.

However, in numerous other cases, the initial link is no more perceivable by speakers. For instance, *ru gereftan* 'to become cheeky' (Lit. 'face take') and *dast andâxtan* 'to mock' (Lit. 'hand throw') constitute opaque sequences in synchrony.

CPs feed lexeme formation rules. The fact that N-V combinations serve as inputs to further lexeme formation rules has been noted in several studies (cf. Introduction) and has been considered by some of them as an argument to support the "wordhood" of these sequences. For instance, the suffix *-i* forms abilitative adjectives from verbs, e.g. *xordan* 'eat' > *xordani* 'edible' (and by further conversion > *xordani* 'food'). This suffix combines with CPs, independently of whether they are compositional or not: *dust daštan* 'to love' > *dustdaštani* 'lovely'; *xat xordan* 'to be scratched' > *xatxordani* 'scratchable'; *juš xordan* 'to bind' > *jušxordani* 'linkable'.

(Non-)predictibility of the verb. Finally, the combination of a particular verb with a particular noun is idiosyncratic in the sense that there is sometimes no semantic justification for the choice of a particular verb. Thus, two semantically close or even synonymous nouns can be combined with two different verbs to give rise to almost synonymous CPs: *hesâdat kardan* (Lit. 'jealousy do') vs. *rašk bordan* (Lit. 'jealousy take') both mean 'to envy', 'to be jealous'; *sohbat kardan* (Lit. 'talk do') vs. *harf zadan* (Lit. 'talk hit') both mean 'to talk', 'to speak'.

3 Productivity of Persian CPs

Although Persian CPs are idiomatic, they are also highly productive. Several theoretical studies have suggested that compositionality is the key to this productivity and put forward hypotheses on how the contribution of the verb and the noun must be combined to obtain the meaning of the predicate (Folli et al., 2005; Megerdoomian, 2012). However, as (Samvelian, 2012) extensively argues, these "radical compositional" accounts are doomed, because they wrongly assume that a given verb and a given noun each have a consistent contribution through all their combinations to form a CP. In this study, we assume that:

1. Persian CPs do not constitute a homogenous class, ranging from fully compositional combinations to fully idiomatic phrases.

2. Compositionality and productivity constitute two distinct dimensions and thus productivity does not necessarily follow from compositionality.

3. A part of Persian CPs can receive a compositional account, provided compositonality is defined *a posteriori*. For these cases, compositionality does account for productivity.

4. For some other cases, analogical extension on the basis of the properties of the whole CP is responsible for productivity.

3.1 Compositionality-Based Productivity

With respect to their compositionality, Persian CPs are comparable to Idiomatically Combining Expressions (Nunberg et al., 1994), idioms whose parts carry identifiable parts of their idiomatic meanings (p. 496). In other words, the verb and the non-verbal element of a CP can be assigned a meaning in the context of their combination. Thus, the CP is compositional (or decompositional), in the sense that the meaning of the CP can be distributed to its components, and yet it is idiomatic, in the sense that the contribution of each member cannot be determined out of the context of its combination with the other one. This is the line of argumentation used by (Nunberg et al., 1994) to support a compositional view of expressions such as *spill the beans*.

Table 1 below illustrates this point. Each line contains a set of CPs formed with *kešidan* 'to pull', where the verb can be assigned a meaning comparable to that of a lexical verb in English.

Examples of CPs with *Kešidan*	
divâr – 'to build a wall', *jâdde* – 'to build a road', *pol* – 'to build a bridge'	> 'build'
lule – 'to set up pipes', *sim* – 'to install cables', *narde* – 'to set up a fence'	> 'set up'
sigâr – 'to smoke a cigarette', *pip* – 'to smoke a pipe', *taryâk* – 'to smoke opium'	>'smoke'
čâqu – 'to brandish a knife', *haftir* – 'to brandish a revolver', *šamšir* – 'to brandish a sword'	>'brandish'
ranj – 'to suffer', *dard* – 'to suffer from pain', *bixâbi* – 'to suffer from insomnia', *setam* – 'to suffer from injustice'	>'suffer from'
dâd – 'to scream', *faryâd* – 'to scream', *arbade* – 'to yell'	>'emit'
harf – 'to extort information', *e'terâf* – 'to extort a confession', *eqrâr* – 'to extort a confession'	>'extort'

Table 1: Meanings of *kešidan* in the context of its CPs

Given that *kešidan* alone cannot convey any of these meanings, these combinations can be considered as ICEs. On the basis of the meaning assigned to *kešidan* and the meaning of the CP as a whole,

new combinations can be produced and interpreted. For instance, the newly coined *šabake kešidan* ‘to install a network’ can be interpreted given the CP *kâbl kešidan* ‘to install cables’ in Table 1.

3.2 Analogical Productivity

CPs such as *šâne kešidan* ‘to comb’, *kise kešidan* ‘to rub with an exfoliating glove’, *jâru kešidan* ‘to broom’ and *bros kešidan* ‘to brush’ constitute a rather coherent paradigm. They all denote an action carried out using an instrument in its conventional way. However, it is impossible to assign a lexical meaning to *kešidan*. Indeed, *kešidan* does not mean ‘to use’, but to use in a specific manner, which cannot be defined without resorting to the noun *kešidan* combines with. Nevertheless, the fact that these instrumental CPs exist enables speakers to create CPs such as *sešuâr kešidan* ‘to do a brushing’ (Lit. ‘hairdryer pull’) on an analogical basis.

In the same way, CPs such as *telefon zadan* ‘to phone’ (Lit. ‘phone hit’), *telegrâf zadan* ‘to send a telegraph’ (Lit. ‘telegraph hit’), *bisim zadan* ‘to walkie-talkie’, ‘to communicate by means of a walkie-talkie’ (Lit. ‘walkie-talkie hit’) constitute a rather coherent paradigm. However, it is impossible to assign a meaning to *zadan* in these combinations. Nevertheless recent combinations such as *imeyl zadan* ‘to email’ or *esemes zadan* ‘to text, to sms’ have been created by analogical extension.

4 A Construction-Based Approach

Building on the conclusions presented in the previous section, Samvelian (2012) proposes a Construction-based approach of Persian CPs. A Construction, in the sense of Goldberg (1995) and Kay and Fillmore (1999), is a conventional association between a form and a meaning. Given that Persian CPs are MWEs, they each correspond to a Construction. Constructions can be of various levels of abstractness and can be organized hierarchically, going from the most specific ones (in our case a given CP, *jâru zadan* ‘to broom’) to more abstract ones (e.g. Instrumental CPs).

Samvelian (2012) applies this Contruction-based perspective to the CPs formed with *zadan* ‘to hit’ and provides a set of abstract Constructions grouping these CPs on the basis of their semantic and syntactic similarities.

Although *zadan* is not the most frequent verb[5] in the formation of CPs compared to *kardan* ‘to do’ or *šodan* ‘to become’, it is nevertheless a productive one, in the sense that it regularly forms new CPs: *imeyl zadan* ‘to email’, *lâyk zadan* ‘to like (on Facebook)’, *tredmil zadan* ‘to run on a treadmill’, *epileydi zadan* ‘to use an epilator’. Besides, *zadan* has a more consistent semantic content than *kardan* ‘to do’ or *šodan* ‘to become’, which function more or less like verbalizers with no real semantic contribution, similarly to conversion or derivation. *Zadan*, on the contrary, can convey several lexical meanings, such as ‘hit’, ‘beat’, ‘cut’, ‘put’, ‘apply’... Consequently, CPs formed with *zadan* provide an interesting case study to highlight the continuum going from lexical verbs to light verbs (or from free syntactic combinations to idiomatic combinations), as well as the way new combinations are coined on the basis of semantic groupings.

Each class is represented by a partially fixed Construction. Here are two examples of Constructions:

(10) **Instrumental-*zadan* Construction**

N0	(*be*) N1	N	*zadan*
Agent	Patient	Instrument	

‘N0 accomplishes the typical action for which N is used (on N1)’

N *zadan*: *bil* – ‘to shovel’, *boros* – ‘to brush’, *jâru* – ‘to broom’, *mesvâk* – ‘to brush one’s teeth’, *otu* – ‘to iron’, *šâne* – ‘to comb’, *sohân* – ‘to file’, *suzan* – ‘to sew’, *qeyči* – ‘to cut with scissors’...

(11) **Forming-*zadan* Construction**

N0	N	*zadan*
Location/Theme	Theme	

‘N is formed on N0’/ ‘N0 is changed into N’

N *zadan*: *javâne* – ‘to bud’, *juš* – ‘to sprout’, *kapak* – ‘to go moldy’, *šabnam* – ‘to dew’, *šokufe* – ‘to bloom’, *tabxâl* – ‘to develop coldsore’, *tâval* – ‘to

[5]To give a rough approximation, the most frequent verb in the Bijankhan corpus (see section 5.1) is *kardan* with 30k occurrences, *zadan* stands in 21st place with 1k occurrences

blister', *yax* – 'to freeze', *zang* – 'to rust', *pine* – 'to become calloused', *nam* – 'to dampen'...

Note that these semantic groupings do not exclusively lie on the semantic relatedness of the nouns occurring in the CPs, but involve the Construction as a whole. While semantic relatedness of the nouns is indeed a good cue for grouping CPs, it does not always allow to account for the relatedness of otherwise clearly related CPs. For instance, *kapak zadan* 'go moldy' (Lit. 'mold hit'), *javâne zadan* 'bud' (Lit. 'bud hit'), *juš zadan* 'sprout' (Lit. 'spot hit'), *šabnam zadan* 'dew' (Lit. 'dew hit'), *zang zadan* 'rust' (Lit. 'rust hit') can be grouped together (see 11 above) on the basis of the fact that they all denote a change of state generally resulting in the formation, development or outbreak of an entity (denoted by the nominal element of the CP) on another entity (denoted by the grammatical subject of the CP). However *mold*, *bud*, *spot*, *dew* and *rust*, *ice*, *dampness* and *blister* do not form a natural class.

Constructions can be structured in networks, reflecting different relationships such as hyponymy/hyperonymy (subtypes vs supertypes), synonymy, valency alternations.

Semantic Subtypes and Supertypes. Some semantic classes can be grouped together into a more abstract class. In this case, the Construction that is associated to them is the subtype of a less specific Construction. For instance the CPs associated to the *Spreading-zadan Construction*, e.g. *rang zadan* 'to paint' (Lit. 'paint hit'), can be considered as *Locatum* (or *Figure*) CPs. *Locatum* verbs, e.g. *paint, salt* (Clark and Clark, 1979), incorporate a Figure (i.e. the noun to which the verb is morphologically related) and have a Ground argument realized as an NP or a PP: 'to paint sth' = 'to put paint (= Figure) on sth (= Ground). In the case of Persian *Locatum* CPs, the Figure is the nominal element of the CP.

Apart from the *Spreading-zadan Construction*, *Locatum-zadan Construction* has several other subtypes: *Incorporation-zadan Construction*, e.g. *namak zadan* 'to salt' (Lit. 'salt hit'), *Putting-zadan Construction*, e.g. *dastband zadan* 'to put handcuffs' (Lit. 'handcuff hit') and *Wearing-zadan Construction*, e.g. *eynak zadan* 'to wear glasses' (Lit. 'glasses hit').

Synonymous constructions. The same Construction can be realized by different verbs, e.g. *kardan* 'to do' and *kešidan* 'to pull' also form Instrumental predicates, e.g. *jâru kardan* and *jâru kešidan* 'to broom'. So, along with *Instrumental-zadan Construction*, there is also an *Instrumental-kešidan Construction* and an *Instrumental-kardan Construction*. These three partially fixed Constructions are subtypes of a more abstract Construction, with no lexically fixed element, namely *Instrumental Construction*. Synonymy rises when the same noun occurs in the same Construction realized by different verbs.

Valency alternating Constructions. The same Construction can display valency alternations. For instance, in an *Instrumental Construction*, the Agent argument can be mapped to the grammatical subject and the Patient to the grammatical object, in which case we obtain an "Active" *Instrumental Construction*, or the Patient can be mapped to the grammatical subject, which gives rise to a "Passive" *Instrumental Construction*. This valency alternation is often realized by a verb alternation in the CP: *otu zadan* 'to iron' vs. *otu xordan* 'to be ironed (Lit. 'iron collide'); *âtaš zadan* 'to set fire' vs. *âtaš geftan* 'to take fire' (Lit. 'fire take').

For a detailed description of Constructions and their hierarchical organization see Samvelian (2012) and Samvelian and Faghiri (to appear).

5 PersPred's Database Conception

Building on Samvelian (2012), *PersPred 1* inventories the CPs formed with *zadan* and a nominal element. Its first delivery includes around 700 combinations grouped in 52 classes and 9 super classes. 22 fields are annotated for each combination.

5.1 Input Data

As Samvelian (2012) extensively argues, the decision whether a given Noun-Verb combination in Persian must be considered as a CP (or LVC) or a free Object-Verb is not straightforward and this opposition is better conceived of in terms of a continuum with a great number of verbs functioning as semi-lexical or semi-light verbs. Consequently, a combination such as *namak zadan* 'to salt' (Lit. 'salt hit') can be viewed either as a CP or as the combination of a lexical verb – *zadan* meaning 'to put', 'to add'

or 'to incorporate' – and its object. Hence, the existence of *felfel zadan* 'to pepper', *zarčube zadan* 'to add tumeric' and many others, which constitute an open class. So, our main concern in the elaboration of *PersPred* is not to solve this insolvable problem. We rather intend to provide a sufficiently rich description of the totally idiomatic combinations as well as semi-productive and even totally productive ones, allowing a precise characterization of the lexical semantics of the simplex verbs in Persian. We thus aim to ultimately elaborate a comprehensive verbal lexicon for Persian.

PersPred is built up, and continues to be enriched, from different types of resources and through complementary methods, in a permanent back-and-forth movement.

1) A first list was established on the basis of Samvelian (2012), which proposes a manually extracted list of CPs from various lexicographic resources, literature, media and the Web, along with their semantic classification.

2) This initial list was enriched in two ways, automatic extraction from the Bijankhan corpus[6] and by manually adding semantically related combinations.

Automatic extraction. We used the Bijankhan corpus (Bijankhan, 2004), a freely available corpus of 2.6m tokens, from journalistic texts, annotated for POS. We first lemmatized the verbs (228 types, 185k tokens)[7] and then extracted CP candidates according to the following pattern : N-V or P-N-V, since, as also mentioned by Tamsilipoor et al. (2012), the N-V pattern can be considered to be the prototypical pattern of the CP construction in Persian. Additionally, in order to include prepositional CPs, e.g. *dar nazar gereftan* 'take into account' (Lit. in view take) or *be zamin zadan* 'make fall' (Lit. to ground hit), we also took into account the noun's preceding element if it was a preposition. In total, we extracted a set of 150k combinations (37k types) regardless of the verbal lemma with, as expected, a large number of hapaxes (25k). For *zadan*, we have 1056 combinations of 386 types with 267 hapaxes. It should be noted that low frequency does not imply the irrelevance of the combination since the frequency is corpus-dependent, for instance well established CPs such as *pelk zadan* 'blink', *neq zadan* 'nag', *havâr zadan* 'scream' or *neyrang zadan* 'deceive' have only one occurrence in the corpus. Hence, the manual validation of all the extracted combination types is necessary. To do so, we stored all the candidates in a spreadsheet sorted by descending order of type frequency and manually filtered out irrelevant sequences.

Manual enrichment. Given the existing classes, we considered a set of new candidates to expand each class on the basis of semantic relatedness. We used a simple heuristic – based on Google search results for the exact expression formed by the noun and the verb in its infinitive form – combined with our native speaker intuition to decide whether a candidate should be retained or not. For instance, given the existence of the class labeled `Communicating` with members such as *telefon zadan* 'to phone' or *faks zadan* 'to fax', we considered combinations such as *imeyl zadan* 'to email' and *esemes zadan* 'to SMS', 'to text'.

Note that for totally productive classes (e.g. *Incoporating* class with members such *namak zadan* 'salt' (see above), listing all potential combinations was useless, since the verb selects the noun it combines with in the same way as a lexical verb selects its complements, i.e. via restricting its conceptual class. So, the actual size of a class in *PersPred 1* does not necessarily reflect its real extension.

5.2 Encoded Information

PersPred 1 contains 22 different fields which are conceived to capture different types of lexical, syntactic and semantic information. Tables 2, 3 and 4 below illustrate these fields via the example of the CP *âb zadan* 'wet'. Note that 2 extra fields provide (at least) one attested example in Persian script and its phonetic transcription.

Lemma information. 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.

`CP-Lemma` indicates the lexical identity of the CP. Consequently there are as many lemmas asso-

[6] http://ece.ut.ac.ir/dbrg/bijankhan/

[7] We took the verbal periphrasis into account in the way that a complex conjugation of, for example, three tokens such as *xânde xâhad šod* 'will be read' or two tokens such as *zade ast* 'have hit', are lemmatized and counted as one verb.

Field	Example
Verb	(V in Persian script)
Noun	(N in Persian script)
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

Table 2: Lemma fields for *âb zadan* 'to wet'

ciated to the same combination as meanings. Thus CP-Lemma allows to distinguish homonymous CPs on the one hand and to group polysemous and syntactically alternating CPs on the other hand. The notation used is as follows: The CP-lemma is encoded by the concatenation of the nominal and the verbal element, linked by a hyphen and followed by a number, beginning from 0. Homonymous CPs are formed with the same components but refer to clearly different events or situations. For instance, *suzan zadan* (Lit. needle hit) means either to sew or to give an injection. A different lemma is associated to each meaning in this case, suzan-zadan0 and suzan-zadan1. We have adopted an approach favoring grouping of polysemous CPs, by assigning the same lemma to polysemous CPs. Polysemy is hence accounted for by creating multiple lexical entries.

Subcategorization and syntactic information. 8 fields represent the syntactic construction of the CP and its English equivalent through an abstract syntactic template inspired, as mentioned above, by Gross (1975). Valency alternations and synonymy are also represented through 3 fields, Intransitive, Transitive and Synonymous Variants.

The subcatgorization frame is provided by Synt-Construction combined with PRED-N, Prep-Form-N1, Prep-Form-N2, 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 1st argument (subject); N1 the direct object; Prep N1 the prepositional object and so on.

The nominal element of the CP, indicated by PRED-N, is also assigned a number. Even though, this element does not display the typical semantic properties of an argument, from a syntactic point of view it can undergo different operations, which means that it has a syntactic function and must thus be taken into account in the obliqueness hierarchy. PRED-N specifies which constituent in Synt-Construction is the nominal element of the CP (i.e. forms a CP with the verb), and thus takes as its value either N0, N1, N2 or N3 or Prep Nx, in case the nominal of the CP is introduced by a preposition. Prep-Form-N1 and Prep-Form-N2 indicate either the lemma of the preposition which introduces N1 and N2, in case the preposition is lexically fixed, or its semantic value:

Field	Example
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

Table 3: Syntactic fields for *âb zadan* 'to wet'

Alternations in the argument realization (i.e. direct *vs* prepositional) give rise to several entries. For instance, the second argument of *âb zadan* 'to wet', can either be realized as an NP or a PP (i.e. Dative shift alternation). Consequently, *âb zadan* has two entries which differ with respect to their Synt-Construction feature value: N0 Prep N1 N2 V vs N0 N1 N2 V. Note that these two entries are considered to be two different realizations of the same lemma (i.e. they have the same value for CP-Lemma).

Construction-EN-Trans simultaneously provides the English translation of the CP and the way the arguments of the Persian CP (as encoded in Synt-Construction) are mapped with the grammatical functions in the English translation.

Intrans-Variant, Trans-Variant and Syn-Variant provide information about valency alternations and synonymy. 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.

Semantic information. 5 fields are dedicated to semantic information, e.g. the semantic subtype and supertype and the type of meaning extension (metaphor, metonymy, synecdoche), if applicable.

Field	Example
`Sem-Class`	Spreading
`Sem-Super-Class`	Locatum
`Constant-Sem`	Liquid
`Subject-Sem`	Human
`Meaning-Exension`	NONE

Table 4: Semantic fields for *âb zadan* 'to wet'

`Sem-Class` and `Sem-Super-Class` give the semantic classification of the CP, i.e. the semantic class and the semantic superclass which the CP is a member of (cf. Section 4 for a detailed explanation). The value of `Sem-Class` corresponds to the most specific partially fixed Construction of which the CP is an instance. The value of `Sem-Super-Class` is the less specific Construction of which the CP is an instance. These feature allow for a hierarchical organization of CPs in classes and super-classes, implementing the Construction networks mentioned in Section 4. CPs which do not pertain to any of the classes are nevertheless considered as the only member of the class they represent. All these singleton classes are assigned the value "isolated" for `Sem-Super-Class`.

`Subject-Sem` and `Constant-Sem` give the semantic class of the subject and the nominal element of the CP. Our classification is more fine-grained than the one adopted in Wordnet, but it can easily be converted into a Wordnet-type classification.

`Meaning-Extension` indicates if a CP has undergone semantic drift, mainly metaphor, metonymy or synecdoche. In the case of a metaphoric extension, the concerned CP is linked to the CP from which it is metaphorically driven.

The integration of a given CP into a given class has been decided on the basis of its most salient semantic properties or some of its meaning components. It should be noted that some meaning components cut across the classes identified in *PersPred 1* and consequently, the CPs that display these meaning components can be cross-classified in different classes[8]. At this stage, only one specific class (i.e. Construction) is mentioned for each CP. One of the future developments of *PersPred* will be to include multiple class memberships.

6 Conclusion

In this paper, we presented *PersPred 1*, which inaugurates the elaboration of a large-scale syntactic and semantic database for Persian CPs. *PersPred 1* is dedicated to CPs formed with *zadan* 'to hit'. We plan to extend its coverage by integrating CPs formed with *dâdan* 'to give', *gereftan* 'to take' and *xordan* 'to collide' shortly. Bearing in mind that integrating new verbs will have an impact on the semantic classes and their networks, and given the fact that our main difficulties so far have been the semantic classification and the time-consuming task of manual annotation, we are currently elaborating semi-automatic annotating methods in order to achieve a satisfactory pace in the future development of *PersPred*.

Acknowledgments

This work was supported by the bilateral project *PerGram*, funded by the ANR (France) and the DGfS (Germany) [grant no. MU 2822/3-I] and is related to the work package LR4.1 of the Labex EFL (funded by the ANR/CGI). We would like to thank Gwendoline Fox and the anonymous reviewers for their helpful comments.

[8] See (Levin, 1993) for similar remarks on English verb classes.

References

Mohammad Bijankhan. 2004. The role of the corpus in writing a grammar : An introduction to a software. *Iranian Journal of Linguistics*, 10(2).

Olivier Bonami and Pollet Samvelian. 2010. Persian complex predicates: Lexeme formation by itself. Paper presented at Septièmes Décembrettes Morphology Conference, Toulouse, December 3.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. 2007. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41:61–89.

Raffaella Folli, Heidi Harley, and Simin Karimi. 2005. Determinants of event type in Persian complex predicates. *Lingua*, 115:1365–1401.

Adele E. Goldberg. 1995. *A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Adele E. Goldberg. 1996. Words by default: Optimizing constraints and the Persian complex predicate. In *Annual Proceedings of the Berkeley Linguistic Society 22*, pages 132–146. Berkeley.

Adele E. Goldberg. 2003. Words by default: The Persian complex predicate construction. In E. Francis and L. Michaelis, editors, *Mismatch: Form-Function Incongruity and the Architecture of Grammar*, pages 117–146. CSLI Publications, Stanford.

Maurice Gross. 1975. *Méthodes en syntaxe : régime des constructions complétives*. Hermann, Paris.

Gholamhossein Karimi-Doostan. 1997. *Light Verb Constructions in Persian*. Ph.D. thesis, University of Essex.

Simin Karimi. 1997. Persian complex verbs: Idiomatic or compositional. *Lexicology*, 3:273–318.

Paul Kay and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language*, 75(1–33).

Parviz Khanlari. 1986. *Tarix-e zabân-e farsi (A History of the Persian Language)*. Editions Nashr-e Now.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago.

Karine Megerdoomian. 2002. *Beyond Words and Phrases: A Unified Theory of Predicate Composition*. Ph.D. thesis, University of Southern California.

Karine Megerdoomian. 2012. The status of the nominal in persian complex predicates. *Natural Language and Linguistic Theory*, 30(1):179–216.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Elizabeth Ritter and Sara Rosen. 1996. Strong and weak predicates: Reducing the lexical burden. *Linguistic Analysis*, 26:1–34.

Ali Ashraf Sadeghi. 1993. On denominative verbs in Persian. In *Farsi Language and the Language of Science*, pages 236–246. University Press, Tehran.

Pollet Samvelian and Pegah Faghiri. to appear. Rethinking compositionality in persian complex predicats. In *Proceedings of the 39th Berkeley Linguistics Society*. Linguistic Society of America, Berkeley.

Pollet Samvelian. 2012. *Grammaire des prédicats complexes. Les constructions nom-verbe*. Lavoisier.

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.

Mohammad-Mehdi Vahedi-Langrudi. 1996. *The syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi*. Ph.D. thesis, University of Ottawa.