SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 32 (2012) 302 – 307

Procedia
Social and Behavioral Sciences

4th International Conference of Cognitive Science (ICCS 2011)

Annotation of grammatical function in the Persian treebank

Ahmad Pouramini[a,*], Elham Moridi[b]

[a]Sirjan University of Technology, Kerman, Iran
[b]Linguistics Department, Fars Science and Research Branch, Islamic Azad Universitys, Fars, Iran

Abstract

In this paper we present and justify methodological principles and syntactic criteria to design an annotation scheme for a Persian Treebank. The advantages of the proposed scheme for annotation of the Persian Treebank will be discussed. At the same time, we present the way that different types of linguistic knowledge (morphological, syntactic and semantic) are encoded in the structures of the schema. We will show how this scheme can account for many of the syntactic constructions that appear to be unique to the Persian language.

Keywords: Natural language processing; corpus annotation; treebank; Persian language; free-word-order languages

Introduction

Today, treebanks constitute an essential resource both to develop natural language processing (NLP) applications and to acquire linguistic knowledge about how a language is used. So far, no serious attempt has been made to create a Persian Treebank, despite some text and speech corpora that have been developed for specific purposes. One of the most comprehensive and adequate of these corpora for the purposes of NPL is the FLDB (Farsi Language Database). FLDB first released by Assi (1997) at the Institute for Humanities and Cultural Studies. The corpus version was updated in 2005, in 1256 character code pages, and named PLDB2 (Persian Language Database). It includes more than 56 million words (Assi, 2005). One advantage of this database is that to each word four linguistic tags are attached at once: phonetic, syntactic, semantic, and lemma (Assi & Hajiabdolhosseini, 2000).

To our knowledge, no representativeness scheme has been applied. The data of this corpus can be employed in building a comprehensive syntactic annotated corpus for a Persian (*treebank*). But the first step in building such treebank is to establish a careful annotation for it. In this paper we present and justify syntactic criteria to design such annotation scheme. In particular, we focus on annotation of grammatical functions and issues concerning the syntactic annotation of Persian language. We do not aim at the application of one or another linguistic theory, but propose an annotation scheme, neutral enough to be used for any research about Persian and easy to translate into other formalisms.

* Corresponding author. Tel.: +98-913-392-3475; Fax: +98-345-423-9401
E-mail address: pouramini@gmail.com

1877-0428 doi:10.1016/j.sbspro.2012.01.044

1. The Persian language properties

Persian is a member of the Indo-European language family and has many features in common with the other languages in this family in terms of morphology, syntax, phonology, and lexicon. Middle Persian had become more analytical, having no grammatical gender and few case markings, and Persian has inherited such characteristics.

Persian is a null-subject language with a basic SOV word order. The main clause precedes a subordinate clause. The language uses prepositions, uncommon to many SOV languages. The one case marker, *rā*, follows the accusative noun phrase. Normal sentences are structured subject-preposition-object-verb. If the object is specific, then the order is "(S) (O + "*rā*") (PP) V" (Karimi, 2003).

Although it is assumed that the Persian clause has an underlying order, there is fairly free order among constituents at the surface. This is because the parts of speech are generally unambiguous, and prepositions and the accusative marker help disambiguate the case of a given noun phrase. Adverbs appear virtually everywhere within a phrase. These characteristic have allowed Persian a high degree of flexibility for versification and rhyming. However, in formal usage the subject mood is widely used (Ghomeshi, 1996).

Persian allows scrambling. The study shows that scrambling in Persian is mainly the feature of spoken language (Karimi, 2003). It has different kinds; short distance, long distance, multiple and rightward. In noun phrases, the sequence of words is around at least one noun, namely the head word. So, the noun phrase could be either a single unit noun, or a sequence of other elements with a noun. To make a phrase, there are some restrictions for the elements surrounding a head to make a constituent; otherwise the sequence of elements will be ill-formed, that is, ungrammatical. Adjectives mostly follow the noun they modify but there are some compounds in which adjective precede the nouns.

Verbs are inflected in the language and they indicate tense and aspect, and agree with subject in person and number. Compound verbs are very common in Persian. Light verbs such as *kærdæn* (to do, to make) are often used with nouns to form what is called a compound verb, light verb construction, or complex predicate (Mohammad & Karimi, 1992). Some more facts about the language are provided in the remaining sections as needed.

2. Treebank annotation

Different languages encode grammatical relations in different ways and through different morphological and syntactic devices. Typical synthetic expressions of grammatical functions can be found in Latin and case-based languages, while in other languages those functions can be analytically represented through Prepositional Phrases or inflectional morphemes. For instance, in Latin the direct object is in accusative case and indirect object is usually in dative case, while in Persian the direct object usually introduced by the particle *–râ* and indirect object is introduced by a Preposition (usually (*be*)).

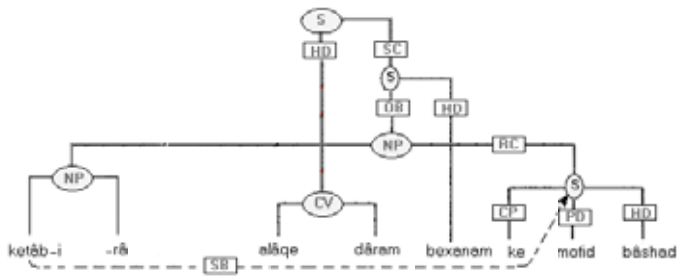


Figure 1: The hybrid structure for the Persian sentence 'ketâb-i râ alâqe dâram bexaram ke mofid bâshad.' (I like to buy a book which is useful.)

Accordingly, the annotation schemes of existing treebanks are classified as constituency-based and dependency-based scheme. A constituency-based annotation scheme makes syntactic constituents (such as noun phrases (NP), verb phrases (VP) and prepositional phrases (PP)) explicit and organizes the sentence in hierarchically structured phrases. A dependency-based annotation scheme concerns the relationships between the words in the sentence in terms of heads and dependents and represents the sentence as a dependency tree or graph. The relations in the syntactic structure can be labeled with grammatical relations or other functions that the dependent plays towards the head.

Constituency is usually employed to annotate languages like English in which there is a fixed constituent order. In this case, there is an exact matching between the position of constituents and their functional roles. On the other hand, dependency annotation is claimed to be more suitable if it is free-word-order language (Bosco, 2004; Bemova, Hajic, Hladka, & Panevova, 1999).

Still, there are some free-word-order languages which use a mixed formalism where the sentence is split in syntactic subunits (phrases), but linked by functional or semantic relations, e.g., the NEGRA treebank for German, the Alpino treebank for Dutch and the Lingo Redwood treebank for English.

3. The annotation scheme for a Persian Treebank

Considering the properties of Persian language, we have chosen the last approach to combine the advantages of both dependency structure and phrase structure representations. The reason for this choice is that, while there is fairly free order among the Persian constituents at the surface, the word order within constituents is quite fixed. At this stage, constituent annotation is convenient for Persian as a previous step for the annotation of the argument structure. In this approach argument structure can be represented in terms of unordered trees. The branches of the tree may cross, allowing the encoding of local and non-local dependencies and eliminating the need for traces. A tree meeting these requirements is given in Figure 1.

The structure is similar to the one employed in NEGRA and TIGER treebanks for German (Skut, Brants, & Uszkoreit, 1998; Brants, Dipper, Hansen, Lezius, & Smith, 2002). In this structure nodes can be either words or phrasal labels (S, VP, PP…). Part-of-speech and morpho-syntactic information is encoded on the word level. It has been argued that many of the syntactic constructions that appear to be unique to Persian can be accounted for by the lexical properties of the inflectional morphemes or features involved (Ghomeshi, 1996). This again shows the importance of annotating the morpho-syntactic information.

All elements which form a single constituent are attached to a phrasal node. The phrasal node can convey the morpho-syntactic information of their associated words. Branches of the tree can be labeled by relations: Presence or absence of a particular label named HD distinguishes headed and non-headed structures. For example, the head word of the sentence is specified by HD, but the head word of a noun phrase could be left as undetermined. This is the case of two NPs and the compound verb in Figure 1. Other relations represent grammatical functions in order to make the argument structure explicit.

As we mentioned before, Persian is a free-constituent-order language where case markers help hearers to identify the syntactic roles of constituents. Therefore, it has been decided to use rather flat trees for the phrase structures. Moreover, as theory independence is one of our objectives, a flat structure is considered more neutral. Whenever a theory-specific structure is desired, positional and part-of-speech information of the elements can be employed to recover the structure from these under-specified representations. Furthermore, a simpler annotation seems a better starting point, because it is always possible to add fine grained annotation over a first shallow one.

The direct consequence of this scheme is the uniform representation of local and non-local dependencies and a transparent structure.

3.1. Treatment of selected phenomena

3.1.1. Compound verbs

As mentioned before, compound verbs in Persian comprise a major part of the verbal system. They are formed by a simple verb combined with a noun, adjective, adverb or prepositional phrase. Persian compound verbs cannot be considered a lexical unit since its elements may be separated by a number of elements (Karimi, 1997). For example,

sentence 1 may appear as follows: *alâge be xândan-e ketâb-i dâram ke mofid bâshad*. Therefore, the elements of compound verbs are separated into two sub trees attached to a particular phrasal node labeled with CV.

3.1.2. Null-subject

Persian is a null-subject language in which the grammatical person of the subject is reflected by the inflection of the verb. The frequency of this feature is high; therefore, as the subject of the sentences is recoverable from the verb, we decided not to employ null-element for this purpose.

3.1.3. Control construction

In cases where a phrase or a lexical item can perform multiple functions, as it can be observed in control construction, an additional edge may be drawn from that item to the controller; thus changing the syntactic tree into a graph.

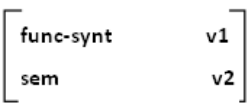
3.2. The annotation of grammatical functions

Due to the rudimentary form of the argument structure representation in our scheme, a great deal of information has to be expressed by grammatical functions. The grammatical functions may carry various kinds of information from purely syntactic functions and thematic roles to functions more proximate to semantics.

Semantic information is currently included in several treebanks. Nevertheless, it is becoming increasingly relevant in NLP tasks, such as Information Extraction or Machine Translation. The semantic features associated to single words looks like the POS tags which instead of morphological information concerns semantic information. The semantic annotation can be employed for the explicit representation of information involved in the predicative structure of the sentence, i.e., the structure that, for instance, a verb forms with its arguments. In the representation of the predicative structure, the structure of the semantic annotation may overlap to structure of the syntactic annotation.

In Persian, most of the predicates in this language are complex predicates and comprise an ever expanding segment of the verbal system (Mohammad & Karimi, 1992). It has been argued in the literature that the argument and event structures of Persian complex predicates, as well as syntactic properties such as control, cannot be simply derived from the lexical specifications of the nonverbal element or the light verb; therefore, suggesting that the syntactic and semantic properties of these elements must be determined post syntactically rather than in the lexicon (Karimi, 1997). This shows that the semantic features of single words are not sufficient for a semantic analysis of Persian sentences. So we are required the annotation of semantic dependencies in the grammatical functions.

Considering above arguments, for the annotation of the grammatical functions in the Persian treebank, we



propose a feature structure including two components as shown in Figure 2:

Figure 2: the feature structure for the annotation of Grammatical Functions in Persian treebank

The functional-syntactic component specifies the dependency type of the words and phrases, and keeps apart arguments and modifiers in the predicative structures. Moreover, this component can make explicit the head of a phrase (label HD). By using the values of this component, the structure distinguishes among a variety of dependency types (e.g., subject, object, indirect object, complement, modifier etc.). At this stage, there should be a trade-off between the granularity of information encoded in the labels and the speed and accuracy of annotation. For example, different arguments of the verb (Subject (SUBJ), Object (OBJ), Indirect Object (INDOBJ), etc.) can be classed as Arguments (ARG). The direct consequence of this method is the availability of another mechanism of under-specification in the annotation or in the analysis of annotated data. Another reason for this choice is that we again prefer to rely on the morpho-syntactic and semantic information to distinguish different type of relations. For example, a noun phrase which is marked with the particle –râ, is the direct object of the sentence.

The semantic component of the structure, which is an optional feature, specifies the role of words and phrases in the syntax-semantics interface and discriminates among different kinds of modifiers and oblique complements (e.g., *be madrese* (to school) in *u be madrese raft* (he went to school) which is the complement of verb *raft* (went) and can be distinguished from other prepositional phrases by a semantic values such as LOC (location))

By following this strategy, the annotation process can be easier, and the result is a direct representation of a complete predicate argument structure, where all the information (morpho-syntactic, functional-syntactic and semantic) is available.

An alternative approach has been followed by the Prague Dependency treebank, which is featured by a three levels annotation (Bemova et al., 1999). This case shows that the major difference between the syntactic (analytic) and the semantic (tectogrammatical) layer consist in the inclusion of empty nodes for recovering forms of deletion. In fact, a part of nodes of the syntactic layer are pruned in the semantic one, e.g., the nodes of an Article and the node of its reference Noun at the syntactic layer are encompassed in a single node at the semantic layer.

As for the annotation scheme we proposed for Persian treebank, since we have access to both words and phrases in the same structure, we are not required to prune some words in order to assign a semantic function to a group of them (phrase). Instead, we simply assign the semantic value to the phrasal node containing these words. Therefore, we are not required a separated semantic layer.

Furthermore, it allows for forms of annotation of relations where not all the features are specified too. In fact, the grammatical functions which specify only the functional-syntactic component allow for the description of syntactic functions which do not correspond with any semantic function, either because they have a void semantic content (e.g., the particle *–râ* or those involved in idiomatic construction) or because they have a different structure from any possible corresponding semantic relation (i.e., there is no semantic relation linking the same linked by the syntactic one). It's especially the case of complex predicates (compound verbs) in Persian in which the light verb and the nonverbal element are separately generated and combined in syntax, and become semantically fused at a different, later level (Mohammad & Karimi, 1992). On one hand, Persian complex predicates cannot be considered a lexical unit since its elements may be separated by a number of elements; on the other hand the meaning of the constructions cannot be obtained by translating each element separately. At this stage, we decided to show each element of the compound verb separately in the structure and use a phrasal category (CV) to distinguish these elements in the sentence. An example of such structure for sentence 1 is given in Figure 3.

Ali hasan-râ davat-e rasmi kard (1)
Ali Hasan+râ invitation+Ezafe formal did
'Ali formally invited Hassan'

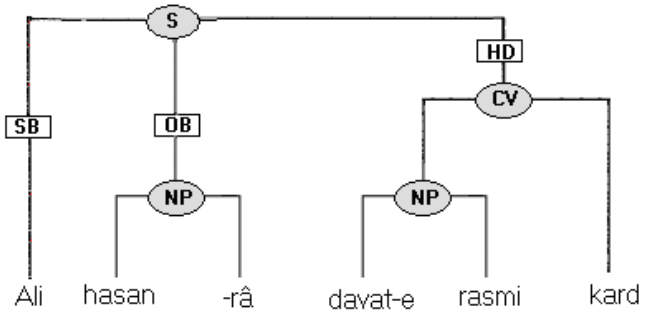


Figure 3. The structure for Persian complex predicates.

In this structure, for each element of the compound verb, the semantic component of its corresponding grammatical function, receives an empty value, while its counterpart in the grammatical function assigned to the whole construction (HD) receives the full meaning of the complex predicate.

A consequence of the hybrid structure we proposed for the annotation of Persian treebank is that there are no problems of inter-layer alignment which must be solved in tasks involving more than one layer (e.g., PP-attachment in parsing which involves both syntactic and semantic knowledge), and which are usually hard to implement because of structural differences among independent levels. For instance, in the Prague treebank which follows the dependency approach for its annotation scheme, the inter-layer alignment is rather complex, because the number of nodes of the semantic level is different from the one at the syntactic level, while in our proposed structure, phrases (group of lexical items) as well as single lexical items can be annotated syntactically and semantically.

4. Conclusion

In this paper we presented an annotation scheme to build a general-purpose treebank for Persian. After examining current annotation approaches, and taking into account the syntactic characteristics of Persian the most appropriate one was selected and its advantages to the annotation of a Persian treebank was discussed. We then presented the treatment of some of the challenging issues in the annotation process such as word order variations and syntax and semantic structure of compound verbs.

Due to the pivotal role of grammatical function in our proposed scheme, we presented a proposal for a careful representation of the information related to the relations in the annotation scheme. We also showed the importance of semantic functions in analyzing Persian complex predicates. In the light of these facts, we proposed a feature structure, which include two components, i.e., functional-syntactic and semantic. By encompassing this linguistic knowledge in the feature structures associated to syntactic units (phrases as well as single words), the annotation scheme features a mono-layered representation of the sentence where the components can make variants of predicative structures explicit in the annotation.

References

Assi, S. M. (1997). Farsi Linguistic Database (FLDB). *International Journal of Lexicography, 10*, 5-7.

Assi, S.M. (2005) PLDB: Persian linguistics database. *Pažuhešgaran (Researchers)*, Institute for Humanities and Cultural Studies Newsletter.

Assi, S. M., & Hajiabdolhosseini, M. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics, 5*, 69-81.

Bemova, A., Hajic, J., Hladka, B., & Panevova, J. (1999). Morphological and syntactic tagging of the prague dependency treebank. *Proceedings of the Association pour le Traitement Automatique des Langues (ATALA) Workshop, France*, 21-29.

Bosco, C. (2004). *A grammatical relation system for Treebank annotation*. Unpublished doctoral dissertation, University of Torino.

Brants, T. S., Dipper, S., Hansen, W., Lezius, W., & Smith, G. (2002). The TIGER treebank. *Proceddings of the Workshop on Treebanks and Linguistic Theories. Sozopol*.

Ghomeshi, J. (1996). *Projection and inflection: A study of Persian phrase structure*. Unpublished doctoral dissertation, University of Toronto.

Karimi, S. (2003). On object positions, specificity, and scrambling in Persian. In S. Karimi (Ed.), *Word order and scrambling*. (pp. 91-124). Oxford: Blackwell Publishing.

Karimi, S. (1997). Persian complex verbs: Idiomatic or compositional. *Lexicology, 3,* 273-318.

Mohammad, J., & Karimi, S. (1992). Light verbs are taking over: Complex verbs in Persian. *Proceedings of The Western Conference on Linguistic (WECOL)*, 195-212.

Skut, W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. *Proceedings of 1st International Conference on Language Resources and Evaluation (LREC'98), Granada*.