An Annotation Scheme for a Persian Treebank

Ahmad Pouramini[1], Naser Mozayani[2]

[1]Department of Computer Engineering,
Iran University of Science and Technology, Tehran, Iran
pouramini@gmail.com

[2]Department of Computer Engineering,
Iran University of Science and Technology, Tehran, Iran
mozayani@iust.ac.ir

Abstract: *In this paper we present and justify methodological principles and syntactic criteria to design an annotation scheme for a Persian Treebank. The main approaches to the annotation of Treebanks are presented in order to account for taken decisions. After examining these approaches, and taking into account the syntactic characteristics of Persian, the most appropriate one will be selected and its advantages 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 proposed schema. We will show how this scheme can provide a useful interface between syntax and semantic.*

Keywords: Natural Language Processing, Treebanks, Annotation Schemes, Free-word-order Languages, Persian Language.

1 Introduction

It is widely admitted that Treebanks constitute a crucial resource both to develop NLP applications and to acquire linguistic knowledge about how a language is used.
So far, there is not a freely available Treebank for Persian, in spite of some references to a large word corpus called *The Farsi Linguistic Database (FLDB*), which comprises a selection of contemporary Modern Persian literature, formal and informal spoken varieties of the language, and a series of dictionary entries [21]. To our knowledge, no representativeness scheme was applied, but some attempts were made to tag the corpus with a POS tagger [22]. The data of this corpus can be employed in building a comprehensive syntactic annotated corpus for 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 methodological principles and 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.

2 Treebank Annotation

2.1 The Annotation Criteria

In order to design an annotation scheme for a Persian Treebank, the annotation criteria of the most significant existing corpora of different languages ([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) have been consulted. The major existing Treebanks are the Penn Treebank developed for English, the Prague Dependency Treebank developed for Czech, and the NEGRA Treebank developed for German. Most of the other Treebanks implements format similar to those of these major Treebanks.
As we have noted, usually Treebank annotation schemes are designed to meet some basic criteria, namely:

Descriptively: Grammatical phenomena are to be described rather than explained.
Data-drivenness: The scheme must provide representational means for all phenomena occurring in texts. Disambiguation is based on human processing skills
Theory-independence: Annotations should not be influenced by theory-specific considerations. Nevertheless, different theory-specific representations shall be recoverable from the annotation.
Following this proposal, we do not wish an application of one or another linguistic theory, but to fix a standard of constituency and functional annotation, neutral enough to be used for any research about Persian and easy to translate into other formalisms.

2.2 The Annotation Scheme

2.2.1 Current Approaches

The annotation schemes of existing Treebanks are generally classified as dependency-based and constituency-based scheme. A constituency-based annotation scheme organizes the sentence in hierarchically structured phrases that span continuous portions of the sentence.
A dependency-based annotation scheme represents the sentence as a dependency tree or graph, i.e. a structure consisting of relations on pairs of syntactic units each composed by a head and a dependent. The relations in the syntactic structure can be labeled with grammatical relations or other specifications of the function that the dependent plays towards the head.
There is an open discussion about the annotation scheme to be assumed when building a Treebank. On one hand, constituency is usually employed to annotate languages like English in which there is a fixed constituent order. in this case, there is an almost exact matching between constituents and functions, that is, the position of a given constituent corresponds to one concrete syntactic function (for instance, in canonical declarative sentences, any noun phrase immediately preceding a verb is usually the subject).
On the other hand, some papers claim that dependency annotation is more suitable if it is free-word-order language [1, 3, 7, 8, 9].
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 [10], the Alpino Treebank for Dutch [9], and the Lingo Redwood Treebank for English [11].

2.2.2 Examining Current Approaches to Persian

Persian is a free word order language that allows scrambling but has basic SOV word order. Although it is assumed that the Persian clause has an underlying order, there is fairly free order among constituents at the surface [13].
Generally for annotation free word order languages, the following features may cause problems:

- Local and nonlocal dependencies from a continuum rather than clear-cut classes of phenomena;
- There exists a rich inventory of discontinuous constituency types (topicalisation, scrambling, clause union, pied piping, extraposition, split NPs and PPs);
- Word order variation is sensitive to many factors, e.g. category, syntactic function, focus;

In the light of these considerations, in the next subsections we examine the issues of adopting each of the mentioned schemes to the Persian annotation.

2.2.2.1 The Consistency-based Alternative

Persian is a free word order language which exhibits word order variation as well as scrambling and discontinues constituents [13]. Serious difficulties can be expected arising from adopting the consistency annotation scheme to the Persian. Due to the frequency of empty categories, word order variation and discontinues constituents in this language (e.g. extraposed relative clauses and separated compound verbs) and scrambling (e.g. extraction out of embedded clauses), the filler-trace mechanism would be used very often, yielding syntactic trees fairly different from the underlying predicate-argument structures.

Consider the Persian sentence below:

(1) ketâb-i râ alâqe dâram bexaram ke mofid bâshad
'I like to buy a book which is useful'

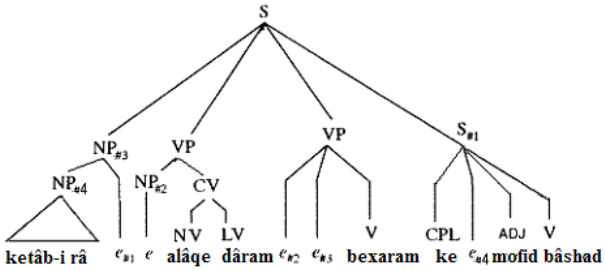

Figure 1: The phrase structure of sentence (1)

The phrase structure of Sentence 1 is given above. The fairly short sentence contains four non-local dependencies, marked by co-references between traces and corresponding nodes. This hybrid representation makes the structure less transparent, and therefore more difficult to annotate.

Apart from this rather technical problem, two further arguments speak against phrase structure as the structural pivot of the annotation scheme for Persian:

Phrase structure models stipulated for Persian differ strongly from each other, presenting a challenge to the intended theory-independence of the scheme. For example several proposals have been made on the phrase structures of the Persian complex predicates [13, 14, 15], impersonal constructions, sentences involving the marker *–râ*, possessive constructions [13], raising constructions, embedded complements [16], the categorical status of so called Persian class 2 prepositions [13], etc.

In Addition, the structural handling of word order variation means stating well-formedness constraints on structures involving many trace-filler dependencies, which has proved tedious.

2.2.2.2 The Dependency-based Alternative

An alternative solution is to make argument structure the main structural component of the formalism. This assumption underlies a growing number of recent syntactic theories which give up the context-free constituent backbone [12].

As described in many theoretical linguistic frameworks, the dependency structure provides a useful interface between syntax and a semantic or conceptual representation of predicate argument structure. For example, Lexical Functional Grammar (LFG) collocates relations at the interface between lexicon and syntax, Relational Grammar (RG) provides a description of the sentence structure exclusively based on relations and syntactic units not structured beyond the string level [1].

While dependency annotation seems to be promising for annotating the Persian Treebank, the sharp distinction between head and dependent stipulated by this formalism causes in general difficulties in the annotation of constructions without a clear syntactic head (e.g., ellipses and coordinations). Moreover, different theories make different headedness predictions. In (2), either a lexical nominalization rule for the adjective *faqir* is stipulated, or the existence of an empty nominal head. Moreover, the so-called DP analysis views the article *ân* as the head of the phrase. In (3), such analysis leads to posit a null definite determiner as the head of the phrase [13]. For another example, there are different approaches to the representation of relative clauses; in some of them, the head of the relative clause is the verb, in others, the head is the relativizer [1].

(2) ân faqir
(3) ketâb-e Hassan

Furthermore, the required theory-independence means that the form of syntactic trees should not reflect theory-specific assumptions, e.g. every syntactic structure has a unique head. Thus, notions such as head should be distinguished at the level of syntactic functions rather than structures.

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.

Finally, non-projective structures can present difficulties for dependency-based as well as for constituency-based frameworks [1].

3 The Annotation Scheme for a Persian Treebank

In light of the pros and cons of each of the approaches, it seems we need a hybrid framework which combines the advantages of dependency structure and phrase structure representations.

Such a hybrid approach has been adopted in the NEGRA Treebank for German and some other treebanks for free word order languages [2, 9, 10, 11]. 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 2.

In this structure nodes can be either words or phrasal labels. Part-of-speech information is encoded in terminal nodes (on the word level). Relations are represented as special nodes in the trees: Headed and non-headed structures are distinguished by the presence or absence of a relation labeled HD, so, it connects a phrasal node (S, VP, PP,..) and its head word, but then again the head of a noun phrase could be left as undetermined. This is the case of two NPs and the compound verb in Figure 2. Then the difference between the particular elements lies in the positional and part-of-speech information, which is also sufficient to recover theory-specific structures from our *underspecified* representations. Other relations represent grammatical functions in order to make the argument structure explicit.

Such a word order independent representation has the advantage of all structural information being encoded in a single data structure and the uniform representation of local and non-local dependencies makes the structure more transparent (Compare Figures 1 and 2).

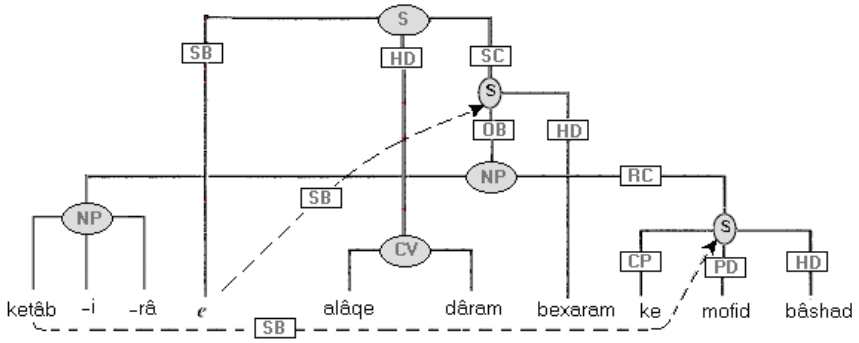

Figure 2: The hybrid structure of sentence (1)

As theory-independence is one of our objectives, the phrase structures are rather flat. For instance it is traditionally said that any sentence has two main constituents: subject and predicate, the second one including the verb, its arguments and its adjuncts. As it is well-known, the relationship between verb and arguments is closer than that between verb and adjuncts. Since Persian is a free word order language, establishing a predicate node could mean having to alter the surface order of the elements in the sentence. Hence, it has been decided not to deal with a predicate constituent. Furthermore a flat structure reduces the potential for attachment ambiguities (such as PP-attachment). Finally, a simpler annotation seems a better starting point, because it is always possible to add new fine grained annotation levels over a first shallow one.

In Figure 2, compound verb *alâqe dâram* is separated into two parts. Note Persian compound verbs cannot be considered a lexical unit since its elements may be separated by a number of elements [14]. For example Sentence 1 may appear as follows: *alâqe be xândan-e ketâb-i dâram ke mofid bâshad.*

Another notable point in Figure 2 is that we treat inflectional morphemes such as *–i* and *-râ* as separated node in the structure. In general, it's the case of all morphemes which can appear after a phrase as well as a lexical item (e.g. [*ketâb-e mofid*]-*i* or [[*ketâb-e mofid*]-*i*]-*râ*).

In order to make explicit the predicate argument structure, in cases of deletion we annotate in the structure, null elements, it's especially the case of pro-dropped subjects (like in Figure 2), which seems to be the main dropped item in Persian language.

In cases where a phrase or a lexical item can perform multiple functions, an additional edge may be drawn from that item to the controller; these additional edges are called *secondary edges* and are represented as dotted lines; thus changing the syntactic tree into a graph.

3.2 The Annotation of Grammatical Functions

Due to the rudimentary character of the argument structure representations, a great deal of reformation has to be expressed by grammatical functions. Their further classification must reflect different kinds of linguistic information: morphology (e.g., case, inflection), category, dependency type (complementation vs. modification), thematic role, etc.

In the next subsection we investigate the major types of knowledge crucial to the Persian syntax,

and present a feature structure for annotating the grammatical functions which allow for a systematic annotation of these types of knowledge.

3.2.1 Grammatical Functions

The argument structure of a sentence is based on the grammatical functions (called *relations*) involved in the sentence. The grammatical functions may carry various kinds of information and the term 'grammatical function' can refer to both purely syntactic functions and thematic roles [5] or functions more proximate to semantics.
Different languages encode grammatical relations in different ways and through different morphological and syntactic devices.
A basic morpho-syntactic distinction concerns the analytic versus synthetic marking of grammatical functions. 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 (which include a Preposition followed by a phrase) or inflectional morphemes (e.g. *–râ* in Persian). 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*)).
Reference [13] argues 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. This shows the importance of annotating the morpho-syntactic information in the grammatical functions of the structure.
Anther major type of linguistic annotation which is currently included in Treebanks is semantic annotation. The semantic annotation can consist in marking of relations or dependencies between words or syntactic units, or marking of semantic features of single words. In practice, the semantic features associated to single words looks like the POS tags which rather than morphological information concerns semantic information, and, in automatic processing, they are often assigned during the POS tagging. The annotation of semantic dependencies can require a separate stage of processing. The marking of semantic relations has been scarcely applied to corpora and only by hand. Nevertheless, it is becoming increasingly relevant in NLP tasks, such as Information Extraction or Machine Translation. The semantic annotation allows, in fact, for the explicit representation of information strictly 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.
As for Persian, most of the predicates in this language are complex predicates and comprise an ever expanding segment of the verbal system [15]; 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 [15]. 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, we propose a feature structure including three components as shown in Figure 3, for the annotation of the grammatical functions in the Persian Treebank:

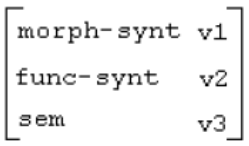

Figure 3:the structure for Grammatical Functions

3.2.2 Morpho-Syntactic Component

The morpho-syntactic component of this structure describes morpho-syntactic features of words or phrases involved in the sentence. This component makes differences between the syntactic behaviors of words or phrases of different categories instantiated by inflectional morphology, explicit. For example the value of this component for a noun phrase which ends with the article *–i* can be *indefinite* (INDEF). Another interesting example is a noun phrase marked with the particle *–râ*, which is called specific direct object in the literature, in such cases, value *definite/specific* (DEF/SPEC) can be assigned to the morpho-syntactic component of the phrase; therefore distinguishing it from nonspecific direct object which have different syntactic behavior in the Persian syntax. It has been argued that the specific direct object appears in a higher position, preceding the indirect object,

while the nonspecific direct object is adjacent to the verb, following the indirect object [18]. The values of the morpho-syntactic component can also be used to distinguish those phrases behaving in ways not conforming to the canonical use of their labeled type (e.g. an NP which plays the role of Adverbial like *emruz sobh* ('this morning') in *emruz sobh u raft* ('he went this morning') In this case a value such as ADV (adverbial) for the NP can be helpful).

3.2.3 Functional-Syntactic Component

The functional-syntactic component identifies 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. One solution is to organize values in a hierarchy according to their different degree of specification (like in [1]). For example in the hierarchy of dependencies, Subject (SUBJ), Object (OBJ), Indirect Object (INDOBJ), etc. can be classed as Arguments (ARG). The direct consequence of this hierarchical organization is the availability of another mechanism of underspecification in the annotation or in the analysis of annotated data. In fact, by referring to the hierarchy we can both annotate and analyze relations at various degrees of specificity.

3.2.4 Semantic Component

Finally, the semantic component of the structure 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* in *u be madrese raft* which is the complement of verb *raft* and can be distinguished form 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 immediately available.

An alternative approach has been followed by the Prague Dependency Treebank, which is featured by a three levels annotation. 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 [4, 5]. In fact, a part of nodes of the analytical layer are pruned in the tectogrammatical one, e.g. the nodes of an Article and the node of its reference Noun at the analytical layer are encompassed in a single node at the tectogrammatical layer.
As for the annotation scheme we proposed for Persian Treebank, since we 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.
The tripartite structure of the grammatical functions guarantees that different components can be accessed separately and analyzed independently (like in [4] or in [18]). 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 a part of components 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 non verbal element are separately generated and combined in syntax, and become semantically fused at a different, later level [15]. 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 can not 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 (like in [15]). An example of such structure for sentence 1, is given in Figure 4.

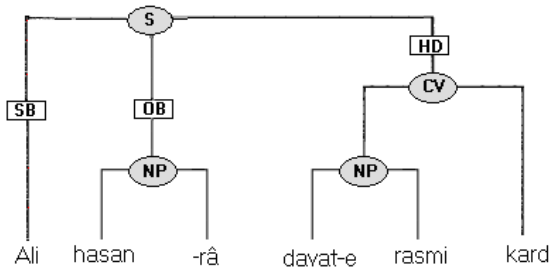

Figure 4: the structure for Persian complex predicates

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

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 tectogrammatical (semantic) level is different from the one at the analytical (syntactic) level; as mentioned before, a part of nodes of the analytical layer which can form a single semantic unit are pruned in the tectogrammatical one, while in our proposed structure, phrases (group of lexical items) as well as single lexical items can be annotated syntactically and semantically.

3.3 The Annotation Process

The construction of a treebank is a particularly labor-intensive and time-consuming task usually performed by human annotators with the help of software tools. Usually the annotation process occurs in two phases: a fully automated Part Of Speech (POS) tagging and a syntactic annotation that can be performed in different ways.
The corpus of texts we are working on in the treebank project is *The Farsi Linguistic Database (FLDB*) [21], and we employed the POS Tagger described in [22] to tag the corpus. Although the main burden of annotation process was and is still carried out by human annotators, but due to similarity between our proposal and the annotation scheme of NEGRA Treebank, the annotation process can follows the approach of the NEGRA project in using an interactive parser.
In this approach an annotator interacts with the parser on POS tagged sentences and a graphical interface displays the structure on the screen for the annotator's decision. The syntactic representation is built incrementally: for each word from right to left, the parser, on the basis of the current grammar, incorporates the current input word in a partial, but fully connected, tree that the user can accept or reject. If the annotator accepts the proposed structure, the parser continues the processing with the next input word, otherwise the parser suggests alternative structures.
Since the project has been recently funded, the annotation process is at early stage and we are working on the tools described above, in order to adopt them for annotating Persian sentences.

3 Conclusions

In this paper we have presented the main criteria to build a Treebank of Persian. Basic methodological principles as well as general syntactic annotation criteria have been presented. In order to find an appropriate annotation scheme, the major approaches have been reviewed and the annotation criteria of the most significant existing corpora have been consulted.
After examining these annotation schemes, and taking into account the syntactic characteristics of Persian the most appropriate one has been selected and its advantages have been discussed. As the selected annotation scheme focuses on annotating argument structure rather than constituent trees, the following features of it are then of particular importance:

- Absence of trace nodes.
- Transparent representation of local and non-local dependencies
- Encoding both phrase-structural information and information on dependency relations
- Being theory independent, and annotating only the common minimum.

In general, the resulting interpreted data also are more neutral with respect to particular syntactic theories which in the case of Persian differ strongly from each other and can pose challenging problems to the annotator.
In the rest of the paper we have been present a proposal for a systematic and careful

representation of the information related to the grammatical functions in the annotation scheme of Persian treebank. First, we have investigated the major kind of linguistic information which is crucial to the Persian syntax. We have mentioned 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. Also we have showed the importance of semantic functions in NLP tasks in general and in analyzing Persian complex predicates in particular.

In the light of these facts, we have proposed a feature structure, which include three components, i.e. morpho-syntactic, functional-syntactic and semantic. By encompassing this linguistic knowledge in the feature structures associated to syntactic units and relations, the annotation scheme features a mono-layered representation of the sentence where the three components can make variants of predicative structures explicit in the annotation.

We have argued that as a consequence of our proposals for the annotation of Persian treebank, we have not problems of inter-layer alignment which must be solved in tasks involving more than one layer (e.g. syntactic and semantic). In general, the resulting interpreted data also are closer to semantic annotation and provide a useful interface between syntax and semantic.

References

[1] Bosco. C. *A grammatical relation system for Treebank annotation*. Ph.D. thesis, University of Torino, 2004.

[2] Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. "The TIGER Treebank," In Proceedings of the Workshop on Treebanks and Linguistic Theories", Sozopol Bulgaria, 2002.

[3] Wojciech S., Brigitte K., Thorsten B. and Hans U. "An annotation scheme for free word order languages," In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP 97), Washington, DC, 1997.

[4] Rambow O., Crecwell C., Szekely R., Taber H., and Walker M., *A Dependency Treebank for English.*, In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02), pp. 857–863, Las Palmas de Gran Canaria, Spain, May 2002.

[5] Bemova A., Hajic J., Hladka B., and Panevova J. "Morphological and Syntactic Tagging of The Prague Dependency Treebank," Journées Atala, Corpus annotés pour la syntaxe, Paris, June 1999.

[6] Marcus M., Santorini B., and Marcinkiewicz. M.A. *Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics*, 1993.

[7] Boguslavsky I., Chardin I., Grigorieva S., Grigoriev N., Iomdin L., Kreidlin L., and Frid N.. "Development of a Dependency Treebank for Russian and its possible Applications in NLP", In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02), pages 852–856, Las Palmas de Gran Canaria, Spain, May 2002.

[8] Oflazer K., Say B., Hakkani-Tür D.Z., and Tür.Building G., *Using syntactically annotated corpora*, chapter Building a Turkish Treebank. Language and Speech. Kluwer, Dordrecht, 2001.

[9] Van der Beek L., Bouma G., Malouf R., and van der Noord. G. "The Alpino dependency Treebank," In *Proc. of CLIN 2001*., 2002

[10] Brants T., Skut W., and Uszkoreit. H., *Building and Using syntactically annotated corpora*, chapter Syntactic Annotation of a German Newspaper Corpus. Language and Speech. Kluwer, Dordrecht, 2001

[11] Montemagni S., Barsotti F., Battista M., and Calzolari. N. *Building the Italian syntactic-semantic treebank*. In Abeill´e (Abeill´e, 2003), pp 189–210, 2003.

[12]

[13] Hudson. R. *English Word Grammar*. Basil Blackwell, Oxford and Cambridge, 1990.

[14] Ghomeshi, J., *Projection and Inflection: A Study of Persian Phrase Structure*. Ph.D dissertation, University of Toronto, 1996.

[15] Karimi, S., "Persian Complex Verbs: Idiomatic or Compositional" Lexicology 3:273-318, 1997

[16] Mohammad J. and Karimi S. "Light verbs are taking over: Complex verbs in Persian" Proceedings of WECOL, 1992, pp 195-212.

[17] Darzi A. *Raising in Persian*. ESCOL 93, pp. 81-92, 1993

[18] Karimi S. 'On object positions, specificity, and scrambling in Persian.' In Simin Karimi (ed.) *Word Order and Scrambling*. Oxford: Blackwell Publishing, pp 91-124, 2003.

[19] Darzi A., *Non-finite control in Persian.* Studies in the Linguistics Sciences, pp. 21-32, 2001
[20] Hashemipour M., *Pronominalization and Control in Modern Persian.* PhD dissertation, University of California, San Diego, 1989.
[21] Karimi S., "Is scrambling as strange as we think it is?" In MIT Working Papers in Linguistics 33, ed. Karlos Arregi, Benjamin Bruening, Cornelia Krause, and Vivian Lin, 159-190, 1999
[22] Assi, S. M. "Farsi Linguistic Database (FLDB)," International Journal of Lexicography, 1997, Vol. 10, No. 3, EURALEX Newsletter p. 5.
[23] Assi S. M. and Haji Abdolhosseini M. "Grammatical Tagging of a Persian Corpus" Institute for Humanities and Cultural Studies, Tehran, Iran, 2000