Grammatical Tagging of a Persian Corpus

S. Mostafa ASSI and M. Haji ABDOLHOSSEINI

Institute for Humanities and Cultural Studies
Tehran, Iran

Abstract

The purpose of this article is to briefly introduce an interactive P.O.S. tagging system developed as a project at *the Institute for Humanities and Cultural Studies* in Tehran, Iran. The system is designed as part of the annotation procedure for a Persian corpus called *The Farsi Linguistic Database (FLDB)*[1], and is the first attempt ever to tag a Persian corpus. In section I, the project itself will be introduced, section 2 presents an evaluation of the project and section 3 is allocated to some suggestions for future work.

1. The Project

1.1. Software

Grammatical tagging has been a very active field lately and a lot of work has been done to tag corpora created for a variety of languages[2]. Needless to say, designing a fully automatic tagging system calls for a lot of experience in the field, enough manpower and considerable funds. Since, this project was carried out as a pilot research, and it was in fact the first attempt to do such work, we decided to design an interactive system. The emphasis is on the methods proposed in Schuetze (1995); nevertheless, all the algorithms were arrived at by trial and error.

The method introduced in Schuetze (1995) is based on the hypothesis that *Syntactic behavior is reflected in co-occurrence patterns*. Therefore, the similarity between two words will be measured with respect to their syntactic behavior to, say their left side by the degree to which they share the same neighbors on the left. Schuetze has applied this method to English for the first time, and has proposed that it should be applied to other languages as well, especially those whose morphology is more complex than that of English. Applying this method to Persian would therefore not only contribute to the simplicity of the project at hand, but could also have some significant theoretical consequences[3]. Schuetze has applied his method in four tests to classify the lexical items in Brown Corpus, and has come up with desirable results

The idea is to gather all the neighbors of a word in two vectors called Left Context Vector and Right Context Vector. In order to limit the size of the vectors, only a certain number of the most frequent word types[4] of the corpus are allowed to enter them. In this project, the number can be between 250 and 1000. The vectors will not have more than 50 dimensions; therefore, in the Persian

[1] A project at the *Institute for Humanities and Cultural Studies* in Tehran which comprises a selection of contemporary Modern Persian literature, formal and informal spoken varieties of the language, and a series of dictionary entries and word lists (Assi, 1997:5).

[2] For more information see Ahrengerg and Jöhansson (1988), Anduriz et al (1995), Chanod and Tapanainen (1995), Elworthy (1995), Feldweg (1995) and Picchi (1994)

[3] It is noteworthy that Persian is a pro-drop and relatively free word order language with an inflectional verb system.

[4] Type is the individual examples of different words or combination of words occurring in a given corpus. (R.R.K. Hartman and G. James. *Dictionary of Lexicography*)

tagger, whenever a context vector of a word reaches 50 dimensions, no more neighbors will be added to it. The context vectors for every word type in the text add up to a file. Each record of this file contains one word type from the text together with up to a hundred of the most frequent words that have appeared immediately before and after that word type. Words with a very low frequency, i.e. less than ten in this project, are ignored, because it has been observed that rare words will have empty context vectors.

Afterwards, the word types are categorized according to their distributional similarity (their similarity in terms of sharing the same neighbors), and then each category can be manually tagged. In addition, it is possible to tag a number of the most frequent words (again 250-1000) of the corpus manually. The software can use both of these methods together, meaning it can categorize the manually tagged words, and then using the distributional method, it can add any untagged word to the word classes that have already been induced. In order to obtain a higher degree of accuracy, three single-member classes have been predefined for the program. These classes include *RA* (the so-called direct object marker), *KE* (relative pronoun), and *XOD* (emphatic pronoun), along with ten numerical digits

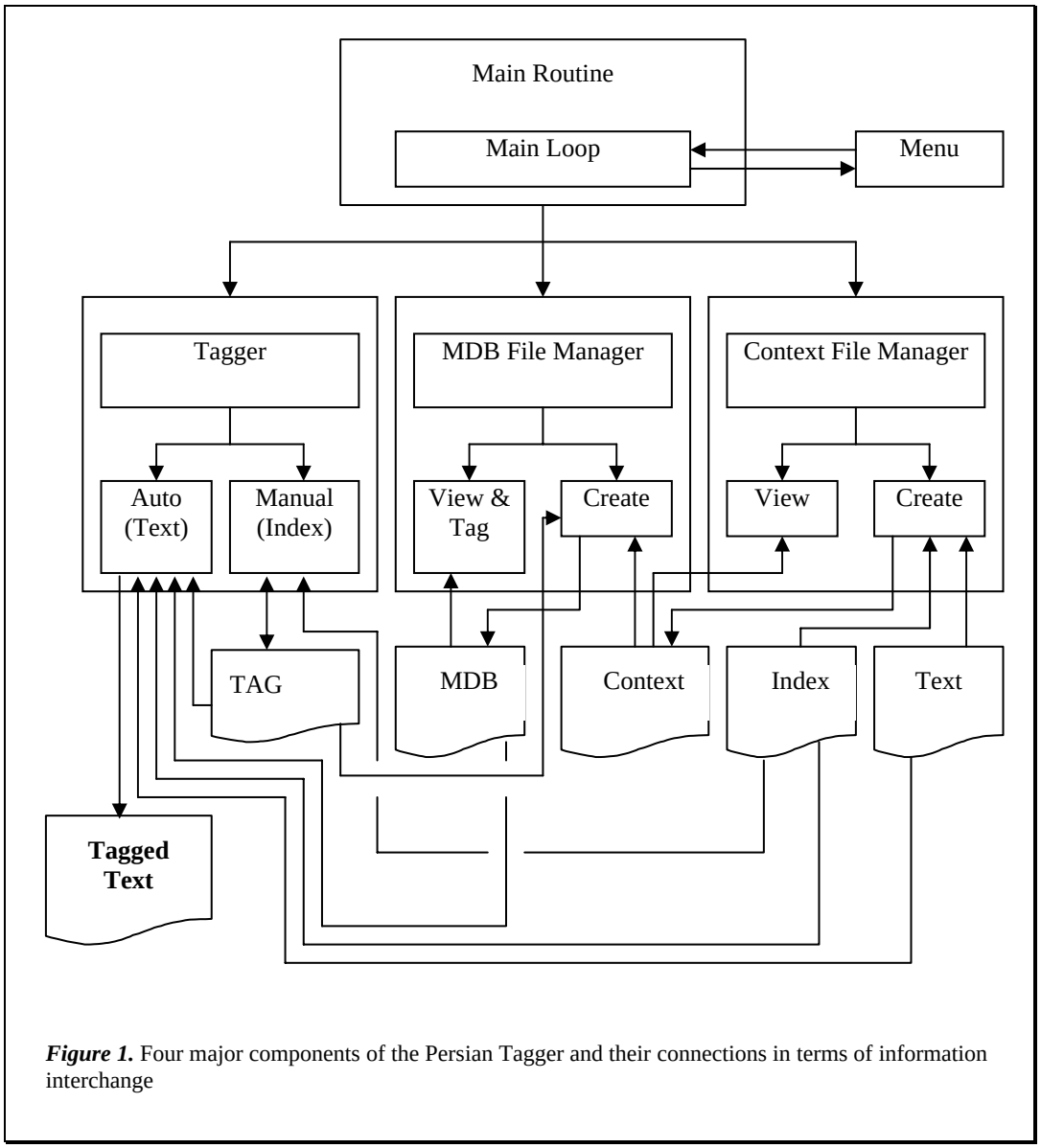

Figure 1. Four major components of the Persian Tagger and their connections in terms of information interchange

(0,1,2…) and punctuation marks.

The Persian Tagger is a program that runs under MS-DOS and requires an IBM or IBM compatible PC. Functionally, the program can be regarded as having four major components:

1. Main Loop, which manages the main menu, error trapping and calling the higher level functions and subroutines;
2. Context File Manager, that creates context vectors and makes their viewing possible;
3. MDB File Manager, that makes possible categorizing the word types recorded in context files, and provides an environment for manually tagging and viewing word categories; and
4. Tagger, which in turn is comprised by two sub-components: (a) a manual tagger, and (b) an automatic tagger. With the manual tagger, it is possible to tag 250-1000 of the most frequent words of a text, while the automatic tagger uses all the data available to it to create a tagged text.

In order for the program to work, an alphabetically sorted frequency word list (called *index* in this project) has to be prepared before hand. This is done using the Oxford Concordance Program (OCP). The frequency word list that the tagger needs is a sequential file containing the word types of the text being tagged arranged in alphabetical order. The file also includes the frequency of each word type together with the line numbers in text, where the tokens[5] of the word type can be found. Figure 1 shows the components of the program and their relationships in terms of information interchange.

The question that might arise here is "Why Scheutze's approach?" As mentioned earlier, this has been the very first attempt to tag a Persian corpus. Therefore, we needed a simple way to break the ice and tackle complications as we step on. Scheutze's approach provided that simplicity for us; furthermore, applying his method on Persian would be a contribution to this new and interestingly simple approach.

1.2. Tagset

The project uses a tagset comprising 43 tags for lexical categories, one tag for single letters that appear in texts as lexical items, and one for unidentified word types. Each tag is made up of a minimum of one and a maximum of five English letters placed between square brackets [ ] inside the texts.

In the design of the tagset, the criteria proposed in Leech (1993: 278-280) were followed. The criteria have been suggested from three different points of view: (a) from the annotator's point of view, (b) from the user's point of view, and (c) some external linguistic criteria.

1.2.1. Criteria that apply to annotation schemes[6]

1.2.1.1. Desiderata from the annotator's point of view

The main three desiderata he proposes here are *speed*, *consistency* and *accuracy*. In order for an annotator's scheme to adhere to these three criteria, it has to be a simple one. This is because a simple scheme would be much easier and more error-free. Besides, if a manual annotation is being carried out, then a higher speed and more consistent results will be arrived at. Furthermore, errors can be detected more easily in a simple scheme. This can result in a higher speed.

1.2.1.2. Desiderata from the user's point of view

From the user's point of view, the following three criteria should be considered in the analysis: *delicacy*, *purpose* and *theory-neutrality*. The annotator's desideratum of speed conflicts with the need of delicacy of analysis, which is often important to the user. In addition, the annotator's objectives for annotation can require a different approach towards it. A corpus may for instance be primarily designed for lexicographical purposes or for more abstract analyses such as syntactic or semantic. Thus, Leech (p. 279) points out that:

> *"It is important in one's general approach to annotation schemes to allow for variable delicacy as one aspect of descriptive reliability of annotation schemes."*

[5] Token is an instance of a graphic word occurring in a given corpus. Frequency counts are typically cited in number of tokens, e.g. 1,000,000 word forms (tokens) in running English text will repeat only about 27,000 different words (types). (see note 4)

[6] This section outlines the criteria proposed in Leech (1993).

Theory-neutrality is set forth for the reason that a scheme which is strongly attached to one theory may satisfy a few but dissatisfy many others. It is therefore, better to assign tags that are useful for a wide range of users. There is a relation between the annotator's condition of simplicity and the user's condition of theory-neutrality. Generally, a simpler scheme is less likely to violate the presumptions of this or that theory.

1.2.1.3. External Linguistic Criteria

The categories recognized in an annotation scheme have to be linguistically real, and not just be means of artificially reducing errors. In the Brown and LOB tagsets, for example, the word *one* was given the unique tag CD1. Whereas numerical *one* should be distinguished from the substitute pronoun *one* (with the plural *ones*) as well as from the indefinite personal pronoun *one* (with the possessive *one's* and the reflexive *oneself*).

In addition, every grammatical tag should be viewed as a complex symbol, each representing a bundle of features. One can design tags in such a way that they reflect these features. For instance, the different forms of the verb *to be* had tags VB0, VBZ, VBD etc. in the CLAWS2 tagset. In this case, V indicates that the word is a verb. B tells us that the verb is *to be*, and the third letters represent the verb's different inflectional forms.

Another matter pointed out by Leech is that in all the annotation schemes today the lexical items are annotated with the assumption that they either belong to a category or they do not. However, experience with corpora suggests that uncertainties of category assignment are quite frequent: not merely because of failures of human understanding, but because of prototypical or fuzzy nature of most linguistic categories. It is therefore, better to try to indicate ambiguity in an annotation scheme. A tag such as ?PNN/NP1 (meaning that the word is either a singular common noun or a singular proper noun) should then be regarded as the honesty of the annotator, rather than his weakness.

1.2.2 Some Points about Persian

Before dealing with the actual tagset, some basic facts about Persian seem to be in order. Persian is an Indo-European language belonging to the Indo-Iranian branch. The grammar of the language has been largely simplified since Old Persian and the vocabulary has been greatly influenced by Arabic and to a less extent by French. Only verbs in this language are inflected. There are fewer tenses in Persian than in English; however, Persian enjoys a widely used subjunctive mood. Persian is an SOV language and direct objects are marked by its only postposition *rä*, which also functions as a topic marker in spoken language. There are disputes over the functions of *rä*. Therefore, we shall resort to the traditional term "direct object marker." However, to show that we do recognize that it has other functions as well we have placed it in the category RA whose only member is the postposition in question. The language does not make use of gender—not even the he/she distinction that exists in English. Only nouns are made plural in this language and even uncountable nouns can be made plural. Possessiveness is indicated through the genitive morpheme *–e*, which is invisible in writing. The existence of a direct object marker enables speakers of Persian to use subjects and objects in a free order although the standard usage is SOV. Adverbs appear virtually everywhere. Adjectives mostly follow the noun they modify but there are some compounds in which adjective precede the nouns. Verb and noun phrases are the most consistent phrases in terms of word order and this enabled us to use more tags that correspond to members of these phrases. Some more facts about the language are provided below as needed.

1.2.3. The Tagset Designed for the Persian Tagger

The method used in this project does not allow for a large tagset because relations among non-adjacent words are not considered in distributional method. Moreover, the inflectional property of Persian verb system requires morphological analyses in order to account for person, number and tense of the verbs. Another problem that we faced in this project was not related to the method but to the problems with Persian orthography. The alphabet system for Persian is an adaptation from Arabic alphabet where, letters are attached to one another to form words. Inconsistencies in what categories should be attached together and what should be written separately are quite common. For instance, the plural form of the word *ketäb* (book) may be written as something like *ketäbhä* or *ketäb hä* (books). Similarly, the equivalent to the phrase "based on this" can be written like *banä bar in*, *banäbar in* or *banäbarin*. In addition, some simple features in Persian grammar lead to problems in tagging. For example, the personal pronouns of Persian have identical forms for subjective and objective cases as well as possessives.

These issues and the fact that this was a pilot project compelled us to keep the scheme as simple as possible. Therefore, we kept the tagset small covering only major word categories especially the ones that were most likely to be recognized through distributional method.

The tagset is made up of 45 tags that have been designed with reference to the categories

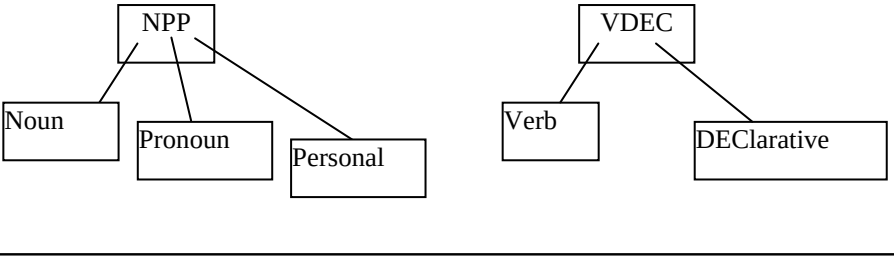

Figure 2. The Internal Structure of Two Tags

normally introduced in dictionaries[7] and to the analysis of surface structure forms of Persian sentences in Meshkatoddini (1994). Each tag is made up of one to five characters. When designing each tag, we tried to preserve the readability of their components, while trying to maintain their fundamental structure. That is, we have tried to allocate a particular character or set of characters to refer to a certain feature, so that ambiguous tags are not formed and no problem arises for an automatic search. In addition, some tags have been set aside especially for ambiguous word types, so that it could be possible to search for and identify them automatically.

There are a few tags in the tagset which include a slash (/) such as NPP/A. This means that the word to which that particular tag is assigned functions as the category named on the left side of the slash, but is not exactly that. NPP/A, for instance, is used for forms such as /*marä*/ (/*man*/ + /*rä*/, i.e. I + ACCUSATIVE marker). In a sentence, /*marä*/ is always located where a personal pronoun (NPP) can appear; nevertheless, it can never be the subject of a sentence. The normal form of a personal pronoun, on the other hand, can be the subject as well as the object in Persian (with /*rä*/ following them as a separate word type). The tag NPP/A for /*marä*/ then shows that the word is a personal pronoun, but it has an accusative marker attached to it. (A complete list of tags is presented in Table 1 of the Appendix.)

The first character or set of characters in each tag represents its general category. The letters that come on the right represent more subtle divisions. Figure 2 below shows the structure of two tags.

2. Evaluation of the Scheme

2.1. Experiment

Experiment with the data showed that our simple adaptation of Scheutz's method worked well with numbers (NUMC), different categories of verbs (e.g. VAUX or VSUB) and nouns (N). Accuracy in these categories was 69-83 percent. However, adverbs and adjectives were the most inconsistent categories. In general, the accuracy of the automatic part of the system proved to be 57.5 percent.

2.2. Problems

The automatic part of the software cannot tag less frequent words of texts; nevertheless, by making amendments in the frequency word list used by the program, it will be possible to tag more words. The first one thousand most frequent word types of a text, however, make up some fifty to seventy percent of the tokens of that text.

Disambiguation is not possible in this system. That is why some tags refer to ambiguous word types, so that they can be searched automatically and disambiguated by another system.

[7] Such categories are significant because of their independence of any particular theory. Thus implementing them in the scheme, one can adhere to the desideratum of theory-neutrality.

The automatic part of the system is quite efficient in categorizing numerals, nouns and verbs; however, when it comes to adjectives, adverbs or some more intricate distinctions such as pronouns vs. nouns, accuracy diminishes substantially.

2.3. Advantages

Experiment on *distributional part of speech tagging* in Persian indicates that the method is also applicable in this Language. The average accuracy obtained is almost the same as the one obtained by Schuetze in a similar experiment. It is, then, plausible to assume that with better computation in line with Schuetze's work better results can be achieved.

Another advantage of the program is its speed. Using the manual section of the tagger, fifty to seventy percent of the tokens in a text can be tagged in a matter of a few hours. The automatic section, on the other hand, can tag a text in just a few minutes. This facilitates testing the accuracy of the system in categorizing word types by using various parameters.

3. Suggestions for Future Work

It seems appropriate to use the tagged texts created by this tagger to train other future tagging systems that use stochastic methods or the Hidden Markov Model. However, for the betterment of results, it is possible to record the number of co-occurrences of word types as well as the tokens themselves. Then in measuring similarities, these values can be included in the calculations as coefficients. If *w* co-occurs with *x* only once in the whole text, then it should not be given as much weight as *y* that appears, say twenty times more next to *x*.

Secondly, one can do what Schuetze has done in his second experiment. That is, rather than recording only word types, it would be much more accurate to record what classes of words a word type can take as its immediate neighbors. This is the very first application of the tagged texts: using the output of the tagger in order to optimize its own performance. As mentioned earlier the tagset used in this project includes tags that represent ambiguous items. A further step can be designing a disambiguation system that searches such items and disambiguates them automatically.

Some relevant activities that have already started in the field include an initiative to arrive at an encoding standard for Persian texts. In addition, as part of the FLDB project, work is in progress to encode Persian homographs and collocations.

References:

Ahrenberg L. and A. Jöhansson. (1988). "An Interactive System for Tagging Dialogues," *Literary and Lingistic Computing*, Vol 3, No. 2, pp. 66-70.

Anduriz, I. et al. (1995). "Different Issues in the Design of a Lemmatizer/Tagger for Basque," From Text to Tags: Issues in Multilingual Language Analysis. On line proceedings of the ACL SIDGAT Workshop at http://xxx.lanl.gov/find/cmp-lg

Assi, S. M. (1997). "Farsi Linguistic Database (FLDB)," *International Journal of Lexicography*. Vol. 10, No. 3, EURALEX Newsletter p. 5.

Chanod, Jean Pierre and Pasi Tapanainen. (1995). "Creating a Tagset Lexicon and Guesser for a French Tagger," From Text to Tags: Issues in Multilingual Language Analysis. On line proceedings of the ACL SIDGAT Workshop at http://xxx.lanl.gov/find/cmp-lg

Elworthy, David. (1995). "Tagset Design and Inflected Languages," From Text to Tags: Issues in Multilingual Language Analysis. On line proceedings of the ACL SIDGAT Workshop at http://xxx.lanl.gov/find/cmp-lg

Feldweg, H. (1995). 'Implementation and Evaluation of a New German HMM Model for POS Disambiguation," From Text to Tags: Issues in Multilingual Language Analysis. On line proceedings of the ACL SIDGAT Workshop at http://xxx.lanl.gov/find/cmp-lg

Leech, Geoffrey. (1993). "Corpus Annotation Schemes," *Literary and Linguistic Computing*. Vol. 8, No. 4, pp. 275-281.

Meshkatoddini, Mehdi. (1994). *An Introduction to Persian Transformational Syntax*. Third Revised Edition. Ferdowsi University Press.

Picchi, Eugenio. (1994). "Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer of Italian," Euralex '94: Proceedings. Papers Submitted to the 6th Euralex International Congress on Lexicography in Amsterdam, the Netherlands.

Schuetze, Hinrich. (1995). “Distributional Part-of-Speech Tagging,” *From Texts to Tags: Issues in Multilingual Language Analysis*. Online Proceedings of the ACL SIDGAT Workshop. On the Internet at http://xxx.lanl.gov/find/cmp-lg

Appendix

Table 1. *The tagset used for Persian Tagger*

No.	Tag	Complete Tag Name	Description	Example
1.	**ADJ**	Adjective	Any word or compound distinctly functioning as an adjective	*bozorg* (big)
2.	**ADJC**	Adjective-Comparative	Comparative adjectives bearing the ending *-tar* (-er)	*bozorg-tar* (bigger)
3.	**ADJN**	Adjective-Noun	Forms ambiguous between adjectives and nouns	*por* (ful) and *par* (feather) have identical spelling
4.	**ADJS**	Adjective-Superlative	Superlative adjectives bearing the ending *-tarin* (-est)	*bozorg-tarin* (biggest)
5.	**ADVI**	Adverb-Interrogative	Equivalent to wh-words in English questioning adverbs	*chetor* (how)
6.	**ADV**	Adverb	Any distinctly recognizable adverb other than those specified in this tagset	*šetäb-än* (hurriedly)
7.	**ADV/C**	Adverb-Complement	Prepositional phrases appearing as a single forms in orthography	*be-to* (to-you) or *baräy-aš* (for-him)
8.	**ADVJ**	Adverb-Adjective	Forms ambiguous between adverbs and adjectives	*xub* (good/well)
9.	**ADVN**	Adverb-Noun	Forms ambiguous between adverbs and nouns	*sar-anjäm* (finally/end)
10.	**ADVP**	Adverb-Place	Adverbs of place	*in-jä* (here)
11.	**ADVPR**	Adverb-Preposition	Forms ambiguous between adverbs and prepositions	*birun* (out/out of)
12.	**ADVT**	Adverb-Time	Adverbs of time	*hälä* (now)
13.	**ATD**	Attribute-Demonstrative	Demonstratives	*in* (this)
14.	**ATD/A**	Attribute-Demonstrative-Accusative	Combination of demonstratives with *RA* the so-called direct object marker	*in-rä* (this-ACCUSATIVE)
15.	**ATD/K**	Attribute + Subordinator	Combination of demonstratives and the subordinator *ke* appearing in a single form in orthography	*än-ke* (corresponding to the relative pronoun who)
16.	**ATE**	Attribute-Exclamation	Exclamations used in the specifier position of noun phrases	*ajab* in *ajab ketäb-i* (what a book!)
17.	**ATI**	Attribute-Interrogative	Question words used in the specifier position of noun phrases	*kodäm* (which)
18.	**ATU**	Attribute-Unspecified	Indefinite articles	*har* (every)
19.	**CONJ**	Conjunction	Any conjunction	*va* (and), *yä* (or)
20.	**N**	Noun	Any distinct noun other than those specified in this tagset	*Ketäb* (book)
21.	**NPP**	Noun-Pronoun-Personal	Personal pronouns These pronouns are used in subject and object position alike in addition to being used as possessive adjectives and pronouns.	*man* (I)
22.	**NPP/A**	Noun-Pronoun-Personal	Combination of NPP and RA (the direct object marker) appearing as one unit in	*to-rä* (you-ACCUSATIVE)

No.	Tag	Complete Tag Name	Description	Example
			writing	
23.	**NPREF**	Noun-Pronoun-Reflexive	Reflexive and emphatic pronouns	*xod-am* (myself)
24.	**NPEM**	Noun-Pronoun-Emphatic	The emphatic form without the ending specifying the person.	*xod* (self)
25.	**NPKE**	Noun-Pronoun-KE	The relative pronoun *ke*	*ke* (that, who…)
26.	**NPU**	Noun-Pronoun-Unspecified	Indefinite pronouns	*hame* (everyone)
27.	**NPREC**	Noun-Pronoun-Reciprocal	Reciprocal pronouns	*hamdigar* (each other)
28.	**NUMC**	Number-Cardinal	Cardinal numbers	*yek* (one)
29.	**NUMC/**	Number-Cardinal-Unspecific	Unspecific numbers	*dah-hä* (tens)
30.	**NUMO**	Number-Ordinal	Ordinal numbers	*avval* (first)
31.	**NV/P**	Noun (Pronoun) + Verb	Combination of personal pronouns and verbs appearing as one unit in orthography	*u-st* (he-is)
32.	**PART**	Past Participle	Past participle forms of verbs	*raft-e* (gone)
33.	**PREP**	Preposition	Unambiguous prepositions	*be* (to)
34.	**PREP/**	Preposition-Conjunction	The form *tä*, which is ambiguous between a preposition and a conjunction	*tä* (until, to, so that)
35.	**PUNC**	Punctuation	Punctuation marks	. , : “ ”
36.	**RA**	Accusative Marker *RA*	The only postposition of standard Persian, the so-called direct object marker *rä*	
37.	**VAUX**	Verb-Auxiliary	Auxiliary verbs	*bäyad* (must)
38.	**VDEC**	Verb-Declarative	Any declarative verbs other than those specified	*gof-t-am* (I said)
39.	**VDECN**	Verb-Declarative-Noun	Ambiguous forms between past tense third person singular declarative verbs and truncated infinitives functioning as nouns	*xar-id* He bought. Shopping
40.	**VINF**	Verb-Infinitive	Infinitive form of verbs	*xar-id-an* to buy
41.	**VLINK**	Verb-Linking	Linking verbs	*ast* (is)
42.	**VIMP**	Verb-Imperative	Imperative forms of verbs	*bo-ro* Go.
43.	**VSUB**	Verb-Subjunctive	Subjunctive forms of verbs	*be-rav-ad* (if) he goes he (must) go
44.	**/LTR**	Letter	Letters or mistyped partial words	
45.	**???**	Unknown	Unknown items	

Authors:

S.Mostafa Assi, reader and the head of Linguistics Department of the Institute for Humanities and Cultural Studies. E-mail: S_M_ASSI@ihcs.ac.ir.

M. Haji Abdolhosseini, MA graduate in Linguistics from the Institute for Humanities and Cultural Studies and researcher at Payame Noor University. E-mail: mhabdolhosseini@yahoo.com