Persian in MULTEXT-East Framework

Behrang QasemiZadeh[1], and Saeed Rahimi[2]

[1] Iran University of Science and Technology, Computer Department, Narmak, Tehran, Iran
QasemiZadeh@digitalclone.net
[2] Tehran University, Faculty of Literature and Humanities, Enqelab, Tehran, Iran
Saeedrahimiavval@yahoo.com

Abstract. Farsi, also known as Persian, is the official language of Iran, Tajikistan and one of the two main languages spoken in Afghanistan. It is an Indo-European agglutinating language, written in Arabic script. This paper presents the first step in creating Farsi basic language resources kit. This Step comprises the specifications for morphosyntactic encoding, which is based on the EAGLES/MULTEXT model and specific resources of MULTEXT-East. This paper introduces the language i.e. Farsi, with an emphasis on its writing system and morphological properties, and its specifications. Two other important issues introduced in this paper are; one, a novel Part of Speech (PoS) categorization and, the other, a unified orthography of Farsi in digital environment. A lexicon and an annotated corpus are under preparation.

1 Introduction

With information and communication technology (ICT) becoming more and more important, the need for language and speech technology also increases. In order for people to use their native language on the computers, a set of basic provisions (such as tools, corpora, and lexicons) is required. There have been numerous attempts to prepare basic language resources kits for the languages, especially, the languages of little or no commercial interest. With respect to Farsi, there is only little work experimented in this field. [1][2]

The MULTEXT-East project[1] was a spin-off of the EU MULTEXT[3] project. It developed standardized language resources for six languages such as Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as English, the 'hub' language of the project. The main results of the project were an annotated multilingual corpus, comprising a speech corpus, a comparable corpus and a parallel corpus, lexical resources, and tool resources for these seven languages. The most useful part of the MULTEXT-East project was the morphosyntactic resources which consist of three layers, listed in order of abstraction as follows [4]:

1. 1984 MSD: the morphosyntactically annotated 1984 corpus, where each word is assigned its context-disambiguated MSD and lemma.

[1] Multilingual Text Tools and Corpora for Eastern and Central European Languages

2. MSD Lexicons: the morphosyntactic lexicons, which contain the full inflectional paradigms of a superset of the lemmas that appear in the 1984 corpus. Each entry gives the word-form, its lemma and MSD.
3. MSD Specs: the morphosyntactic specifications, which set out the grammar of valid morphosyntactic descriptions, MSDs. The specifications determine what, for each language, is a valid MSD and what it means, e.g., Ncms means PoS: Noun, Type: common, Gender: masculine, Number: singular.

MULTEXT-East provides a comprehensive framework for corpus development. Also, there are a lot of resources according to this framework, e.g. 1984 MSD for several languages. On the other hand, 1984 is available in Farsi and Farsi in return can be suited to this framework as we will show in the following. This can save us time, and money, moreover we can benefit from software reuse.

In this paper, we will try to propose an approach to represent an annotation of Farsi written corpora according to MULTEXT-East framework. As a result, we will have a discussion about Farsi specifications; we, then, propose our MSD Specs for Farsi. The rest of this paper is structured as follows: Section 2 introduces Farsi, its grammar, and its writing system. Section 3 explains the MSD specifications for Farsi, based on MULTEXT-East framework and the discussion in section 3. Related works are described in section 4. Finally, Conclusion and future works are discussed in section 5.

2 Farsi Language

Farsi, also known as Persian, is the official language of Iran, Tajikistan and one of the two main languages spoken in Afghanistan. Farsi is a member of the Indo-Iranian family of the Indo-European languages. Farsi has the properties of agglutinative languages. Even though Farsi is an agglutinative language, the fusional features can also be found in it. [5][6] The majority of affixes in Farsi are suffix with limited prefixes as well. There is no infix detected in Farsi.[5][6][7] Detailed morphosyntactic features of Farsi are described in section 2.1.

After the Arab's conquest in 651 A.D., the Persians adopted an extension of unified Arabic script for writing. Since Arabic is a cursive script, the number of possible shapes that letters actually can adopt exceeds the number of these letters [8]. Letters attach to each other to represent a word. Since Arabic is a Semitic language, it is obvious that how letters must be attached to each other to represent a word. In Farsi, however, due to the fact that it is an agglutinative language, there could be ambiguity in what letters should be written attached together or detached. For instance, the plural form of the word *ketäb* (book) may be written as 'کتابها' *ketäbhä* or 'کتاب ها' *ketäb hä* (books). This results in some difficulties in Farsi text analysis as cited in [9][10][11], i.e. tokenization of Farsi e-text since word boundaries are not clear. Also, the fact that short vowels are not written and capitalization is not used will result in ambiguities that impede computational analysis of the texts. In section 2.1, we will propose a standard for Farsi transcription to solve the problems mentioned above.

2.1 Farsi Transcription and Encoding in Digital Environments

Unicode standard version 4.0 reserves the range 0600 to 06FF for Arabic characters. The important design principles observed in the Unicode standard and relevant to the representation of Arabic script are characters not glyphs. As mentioned in the previous section, Arabic letters can have up to four different positional forms depending on their position relative to other letters or spaces. According to the design principle "characters, not glyphs", there is no individual code for each visual form (glyph) that an Arabic character can take in varying contexts but there exists only one code for each actual letter. The correct glyphs to be displayed for a particular sequence of Arabic characters can be determined by an algorithm. In order to display the characters properly, two special characters namely ZERO WIDTH JOINER (0x200D) and ZERO WIDTH NON JOINER (0x200C) are added to the character codes, either before or after them. The use of these special characters after a code means that a ZWJ or a ZWNJ should be added after the character if the character is not followed by a "right-join causing" character, or a "non-joining character" respectively.

The ISIRI 6219:2002 (Information Technology – Farsi Information Interchange and Display Mechanism, using Unicode) [12] has been proposed as the Farsi standard for using Unicode in digital environment. This standard indicates a subset of Arabic character set in Unicode to be used by Farsi users; but it does not specify which letters must be written in a separate or attached form. On the other hand, "Iran's Academy of Farsi Language and Literature", which is a governmental body presiding over the use of the Farsi language, has created an official orthography of the Farsi language, entitled "Dastoor-e Khatt-e Farsi" (Farsi Script Orthography) [13], for the proper representation of texts in the paper based system of writing.

Unfortunately there exists no standard for Farsi orthography in digital environment. For this reason, we have suggested an approach to represent Farsi electronic texts as we have done for 1984 corpus. According to the proposed orthography by the Academy, Farsi affixes must be written attached to their stem. In some cases when the stem ends in a letter which is a "right-join causing character", the affixe must be attached to the stem with a short space character before it. In order to fulfill this , we have used ZWNJ character as the short space. We have also used a character set based on the proposed standard in [12]. In this way, space characters represent unambiguous word boundaries and the orthography of Farsi e-texts remains consistent with the one which is proposed in [13]. Also, this transcription results in Farsi e-texts which are more consistent with the e-texts of other languages. This could be useful when developing parallel corpora of Farsi and other languages.

We should consider that the policy of text encoding, tokenization, orthography, and corpus tagging are in interaction with each other. For example, in Farsi it is possible that a bound morpheme appears detached from its stem with an intervening space; if we assume space as a delimiter in the tokenization process according to the used orthography, either we have to consider a tag for these bound morphemes during corpus tagging or, we have to consider a more complicated tokenization process as it is cited in [11] [9] (Figure1).

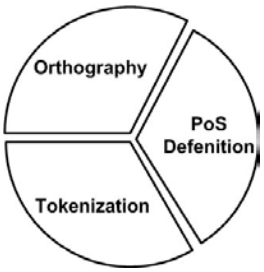

Fig. 1. Consider the whole circle as a proposed standard for corpus tagging in a specific language. Then the tokenization policy, PoS categorization, and language orthography, are fundamental elements that will directly affect the set of tags which is defined for corpus tagging.

2.2 Part of Speech in Farsi

There are seven PoS categories in traditional Farsi grammar [14]: Noun, Adjective, Verb, Adverb, Pronoun, Number, and Interjection. As cited in our previous work [10], this categorization is not adequate enough for analyzing Farsi. We can have more precise categorization considering other aspects of computational analysis of Farsi and comparing it with other languages in multilingual applications. According to our new categorization for PoSs in Farsi, there are 12 categories with their own special attributes. Our concept for this categorization is based both on the position of words in phrasal structure, and also what we have described in figure1. Nouns, Adjectives, Adverbs, Prepositions, Conjunctions, Verbs, Postpositions, Pronouns, Numbers, Determiners, and Interjections comprise our proposed categories. In the following we have discussed salient properties and morphosyntactic attributes of each of these proposed PoSs briefly.

Verbs are usually inflected with number and person. Farsi is neutral for gender. Our categorization divides verbs into 5 major types which are Main, Auxiliary, Copula, Modal, and Light. Most of these types are the same as they are in other languages. The number of Main verbs is limited in Farsi. Modal type of verbs is used to change the aspect of verbs to Subjunctive. Usually they come before Main verbs in present subjunctive form so the Main verb will have normal inflectional attributes. But if the Main verb appears in past 3rd person form, then the construction will be impersonal. Modal verbs usually are not inflected by number and person. However, there is an exception for the verb 'توانستن' (tavânestan) that can be inflected for person and number. Light verbs in Farsi are used to make a compound verb structure. Compound verb structure consists of one or more preverbal elements which could be a noun, adjective, or a prepositional phrase, followed by a Light verb. The number of Light verbs is limited. The elements of a compound verb construction can be separated by other lexical elements such as the object of the verbal construction or an adjective, adverb, etc. Therefore our suggestion is to analyze compound verb construction only at the syntactic level. We should also note that Light verbs are homographic with Main verbs. In Farsi, Past Tense verbs are made using past stem of verbs and present tense is made of present stem of verbs. Future tense is made by the

help of Auxiliary verbs. In order to make progressive form in Farsi, verbs are inflected with the prefix 'می' (mī). Perfective forms of verbs are usually made using auxiliary verbs '... است ،ام' (am, ast, …). Passive form of the verbs in Farsi are made by the help of Auxiliary verbs. Passive form of the verb is made of Past Participle + Auxiliary verb 'شدن' (šodan). In some cases for courtesy, instead of the singular form of the verb, the plural one is used to refer to a singular subject. So we consider it as an attribute for Farsi Verbs. In fact, such attributes for Farsi are not found in traditional grammar books.

In Farsi, Nouns are inflected for number and Definiteness. There is no specific marker, like capitalization in English, for Farsi proper nouns. Plural form is made, similar to English, by adding Plural suffixes to the end of the nouns. Nouns may also be accompanied by the *Ezafe* Marker, a suffix that connects the elements in a phrase, and the indefinite marker. *Ezafe* Marker can appear as 'ی' (ye) when the word ends in certain characters. It can also appear as a short vowel named *Kasre* which sounds "e". In this case, according to the Farsi orthography, it can be deleted from the written text. A noun which is accompanied by *Ezafe* Marker can be considered as the genitive case of the noun.

Farsi adjectives are inflected for degree and definiteness. Adjectives, just the same as nouns, may also be accompanied by the *Ezafe* Marker and in this case we can consider it as a genitive case. Adverbs are often invariable in number. Certain adverbs may appear with the comparative suffix.

In traditional Farsi grammar, the category of determiners is not specified. Considering morphosyntactic specifications of words and the place of them in phrasal structures, we believe that determiners can be specified as a PoS in Farsi. Moreover, this consideration is more consistent with other languages. Most of the words we have considered as determiners here are categorized as adjectives in traditional grammar. There are different types of determiners namely demonstrative, indefinite, interrogative, exclamative, and article. As defined here, there is just one article in Farsi; i.e, 'یک' (yek). It is homonym with 'یک' which is a number.

Farsi has several prepositions but there is only one postposition 'را' (râ). It is an overt marker for direct object. Other categories are almost similar to traditional ones. Detailed description of Farsi grammar can be found in [16] [17] in English and [15] [18] in Farsi.

Table 1. Farsi PoSs and their proper codes according to MULTEXT-East

Part of Speech	Code	Number of Attributes
Noun	N	4
Verb	V	10
Adjective	A	4
Pronoun	P	6
Determiner	D	1
Adverb	R	2
Adposition	S	2
Conjunction	C	2
Numeral	M	1
Interjection	I	0
Abbreviation	Y	0

3 Farsi in MULTEXT-East framework

Table 1 shows the PoSs of Farsi and the number of their attributes. Farsi MSD specification according to MULTEXT-East framework is proposed in table 2. The specification is based on the discussions in section 2. The structure of table 2 is the same as the one in [4]. Table 2 shows the attributes, their position, and the proper values of each proposed PoS. More information about the structure of the table can be found in [19]. We do not show attributes of PoSs irrelevant to Farsi due to their massive volume.

Table 2. MSD specification of Farsi(Farsi)

Nouns

P	ATT	VAL	C
1	Type	common	c
		proper	p
3	Number	singular	s
		plural	p
4	Case	genitive	g
5	Definiteness	no	n
		yes	y

Verbs

P	ATT	VAL	C
1	Type	main	m
		auxiliary	a
		modal	o
		copula	c
		light	l
2	VForm	indicative	i
		subjunctive	s
		imperative	m
		participle	p
3	Tense	present	p
		past	s
4	Person	first	1
		second	2

		third	3
-	----------	----------	-
5	Number	singular	s
		plural	p
-	----------	----------	-
8	Negative	no	n
		yes	y
-	----------	----------	-
10	Clitic	no	n
		yes	y
-	----------	----------	-
14	Aspect	progressive	p
-	----------	----------	-
15	Courtesy	no	n
		yes	y
-	----------	----------	-
16	Transitive	no	n
		yes	y

Adjectives

P	ATT	VAL	C
1	Type	qualificative	f
-	----------	----------	-
2	Degree	positive	p
		comparative	c
		superlative	s
-	----------	----------	-
5	Case	genitive	g
-	----------	----------	-
6	Definiteness	no	n

Pronouns

P	ATT	VAL	C
1	Type	personal	p
		demonstrative	d
		indefinite	i
		interrogative	q
		reflexive	x
		reciprocal	y
-	----------	----------	-
2	Person	first	1
		second	2
		third	3
-	----------	----------	-

4	Number	singular	s
		plural	p
5	Case	genitive	g
		accusative	a
8	Clitic	no	n
		yes	y
12	Animate	no	n
		yes	y

Determiners

P	ATT	VAL	C
1	Type	demonstrative	d
		indefinite	i
		interrogative	q
		exclamative	e
		article	a
		exceptional	x
4	Number	singular	s
		plural	p

Adverbs

P	ATT	VAL	C
2	Degree	positive	p
		comparative	c
7	Case	genitive	g

Adpositions

P	ATT	VAL	C
1	Type	preposition	p
		postposition	t
2	Formation	simple	s
		compound	c

Conjunction

P	ATT	VAL	C
1	Type	coordinating	c
		subordinating	s
2	Formation	simple	s
		compound	c

Numerals

P	ATT	VAL	C
1	Type	cardinal	c
		ordinal	o
		fractal	f
		ordinal2	r
4	Case	genitive	g
6	Definiteness	no	n
		yes	y

Interjection (No Attribute)
Abbreviation (No Attribute)

3 Related Works

Up until now, there are only few works done to create Farsi basic language resources kit. Keyvan et. al. introduces the work done in PersiaNet, a wordnet for Modern Farsi. [20] It has been carried out in an informal setting and is entirely on a volunteer basis. Lexical coverage is currently very sparse. The paper gives a good background about Farsi language and its writing system.

Assi and Haji [21] introduce an interactive PoS tagging system developed as a project at *the Institute for Humanities and Cultural Studies* in Tehran, Iran. It was designed as a part of the annotation procedure for a Farsi corpus called *The Farsi Linguistic Database* [22] *(FLDB)* and is the first attempt ever made to tag a Farsi corpus. The paper mainly emphasizes on the proposed system instead of the morphosyntactic specification of Farsi. The proposed tag set in [21] consists of 45 tags for lexical categories including one tag for single letters that appear in texts as lexical items, and one for unidentified word types. In the tag set, there are some tags that represent ambiguous annotations. As mentioned in [21], it was a pilot project. Studies of the tag set shows that the linguistic backgrounds are based on the traditional approaches and many morphosyntactic features of Farsi have been ignored.

A set of tools for Farsi analysis is introduced in [23]. This project focuses on English-Farsi machine translation. The tools consist of a lexicon and bilingual corpus

[24], and other tools required for the analysis of Farsi. The linguistic background is based on the Machine Translation application and is different from the one proposed here.

5 Conclusion and Future Works

Unfortunately there has not been much effort made to create Farsi language resources. In summary, the significance of this paper can be fallen in two aspects: first, introducing a novel approach to represent Farsi e-text (orthography) in addition to a new PoS categorization. This could be of a great help to solve the problems which are introduced in [20][21][9][10]. Second, introducing a new tag set, according to MULTEXT-East framework, for Farsi corpus tagging.

On the one hand, MULTEXT-East introduces well-established standards with useful tools to manipulate and analyze text corpora. The MULTEXT-East resources are widely available for further researches. On the other hand, Orwell's 1984 which is tagged according to MULTEXT-East for several languages now is available in Farsi. As a result, we have fitted Farsi to this framework. We have started to tag 1984 based on the proposed tag set of Farsi to reach a multi lingual corpus consuming reasonable time and effort.

The production of the corpus and the lexicon is under preparation. Having prepared this corpus, all classical approaches to corpus based linguistics could be applied to Farsi. In this way, we can compare the results with the other efforts that have been done previously on other languages. For example, the prepared corpus can be used for training a tool for automatic PoS tagging of Farsi which uses machine learning techniques. We consider this as a future work.

In order to reach the best results, a novel PoS categorization for Farsi is introduced. One of the most important benefits of this PoS categorization is the consistency it provides for Farsi with other languages without ignoring any information of Farsi grammar. This could be of help during cross linguistic analyses. Also a unified orthography for Farsi e-text is proposed in this paper. The proposed orthography is based on the one proposed in [13] for the paper based system and computational point of view. The proposed orthography is consistent both with other languages and also Farsi grammar. Moreover it is convenient to use.

Acknowledgement

The authors would like to express their sincere gratitude to Prof. Tomaž Erjavec for many fruitful discussions. Also they would like to thank Prof. Damir Ćavar for his constructive idea.

References

1. Strik, H. Daelemans, W. Binnenpoorte, D. Sturm, J. de Vriend, F. and Cucchiarini, C.: Dutch HLT resources: From BLARK to priority lists, In Proceedings of ICSLP, Denver, USA, pp. 1549-1552, Denver, USA, (2002).
2. Krauwer, S. Maegaard, B. Choukri, K. and Damsgaard Jørgensenm, L.: Report on BLARK for Arabic, (2004).
3. Ide N. and Veronis J.: Multext: Multilingual Text Tools And Corpora. In 15th Int. Conference On Computational Linguistics, Pages 588–592, Kyoto, Japan, (1994).
4. Erjavec T., Krstev C., Petkevic V., Simov K., Tadic M., and Vitas D.: The MULTEXT-East Morphosyntactic Specifications For Slavic Languages, Proceedings Of The EACL 2003 Workshop On The Morphological Processing Of Slavic Languages, (2003).
5. Kalbasi, I.: The Derivational Structure of Word In Modern Farsi, ISBN 964-426-128-3, Tehran, (2001).
6. Samare I.: Typological Features Of Farsi, Journal Of Linguistics, Iran University Press, No. 7, pp 61-80, (1990).
7. Keshani, K.: Suffix Derivation in Contemporary Farsi, First Edition, Iran University Press, (1992).
8. Lutz, W.: Unicode and Arabic Script, Workshop "Unicode Und Mehrschriftlichkeit in Katalogen", Sbb Pk, Berlin, (2003).
9. Karine M. And Zajac R.: Processing Farsi Text: Tokenization In The Shiraz Projec,.Nmsu, Crl, Memoranda In Computer And Cognitive Scienc,(2000).
10. Qasemizadeh, B. and Rahimi, S.: Farsi Morphology, 11th Computer Society of Iran Computer Conference, IPM, Tehran, Iran, (2006).
11. Rezaie S.: Tokenizing an Arabic Script Language, Arabic Language Processing: Status And Prospects, Acl/Eacl, (2001).
12. Isiri 6219:2002: Information Technology - Farsi Information Interchange and Display Mechanism, Using Unicode, (2002).
13. Iran's Academy Of Farsi Language and Literature: Official Farsi Orthography, ISBN: 964-7531-13-3, 3rd Edition, (2005).
14. Hasan A. and Ahmadi Givi H.: Farsi Grammar, ISBN964-318-007-7, 22nd Edition, Tehran, (2002).
15. Meshkatodini M.: Introduction to Farsi Transformational Syntax, 2nd Edition, ISBN: 964-6335-80-2, Ferdowsi University Press, (2003).
16. Lazard, G.: A Grammar of Contemporary Farsi, Mazda Publishers, (1992).
17. Riazati D.: Computational Analysis of Farsi Morphology, Msc Thesis, Department Of Computer Science, RMIT, (1997).
18. Bateni, M.: Towsif-E Sakhteman-E Dastury-E Zaban-E Farsi [Description Of The Linguistic Structure Of Farsi Language], Amir Kabir Publishers, Tehran, Iran, (1995).
19. Erjavec T.: MULTEXT-East Morphosyntactic Specifications, Version 3.0. Supported By EU Projects Multext-East, Concede And TELRI, (2004).
20. Keyvan, R. Borjian, H. Kashef, M. and Fellbaum, C.: Developing Farsiet: The Farsi Wordnet, GWC 2006, Proceedings, pp. 315–318, (2005).
21. Assi, S. M. Haji Abdolhosseini, M.: Grammatical Tagging of a Farsi Corpus, International Journal of Corpus Linguistics 5:1, 69–81, (2000).
22. Assi, S. M.: Farsi Linguistic Database (FLDB), International Journal of Lexicography, Vol. 10, No. 3, Euralex Newsletter, (1997).
23. Amtrup, J. W., Mansouri Rad, H. Megerdoomian, K. and Zajac, R.: Farsi-English Machine Translation: An Overview of the Shiraz Project. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-319), (2000).

24. Megerdoomian, K. and Mansouri Rad, H.: Acquisition of Farsi Resources: Corpora and Dictionary Development in the Shiraz Project. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-323).