A Study of Corpus Development for Persian

Masood Ghayoomi[1], Saeedeh Momtazi[2], Mahmood Bijankhan[3]
[1] Freie University of Berlin, German Grammar Group, 14195 Berlin.
[2] University of Saarland, Spoken Language Systems, 66041 Saarbrücken.
[3] University of Tehran, Department of Linguistics, 14155 Tehran.
masood.ghayoomi@fu-berlin.de, saeedeh.momtazi@lsv.uni-saarland.de,
mbjkkhan@ut.ac.ir

Abstract

Persian is one of the Indo-European languages which has borrowed its script from Arabic, a member of Semitic language family. Since Persian and Arabic scripts are so similar, problems arise when we want to process an electronic text. In this paper, some of the common problems faced experimentally in developing a corpus for Persian are discussed. The sources of the problems are the Persian script itself; mixture with the Arabic script; Persian orthography; the typists' typing styles; and mixing Persian code pages with Arabic in the operating systems; linguistic style and creativity in the language.

Keywords

Corpus development, the Persian language

1 Introduction

Text corpus is an electronic source of data to be processed for linguistic investigators or natural language processing applications. A corpus may contain texts in a single language (monolingual corpus) or text data in multiple languages (multilingual corpus). Corpora are the main resource in corpus linguistics to study the language as expressed in samples or real world text. There are several resources in written text that can be used for developing corpora. Newswires and books are the most well-known resources for this task. Nowadays web pages are also widely use as a rich source to construct corpora; because it is possible to collect various texts being a representative of the language by providing the texts in various genres and various authors. The result is providing a reasonable accurate picture of the entire language in which we are interested.

Reaching the goal is not an easy task. While working to build a corpus, you might face difficulties; and before processing the corpus, these problems should be removed at a step named "preprocessing".

In the experiences made to develop a Persian corpus, we faced a lot of problems due to some special features of Persian. In this paper, we discuss these problems to give a comprehensive perspective of Persian corpus development to users who want to work in this area. The structure of the paper is as follows: Section 2 talks about the features that should be kept in mind while constructing a corpus. In Section 3 the developed corpora for

Persian are introduced. Since the paper is about the problems of building Persian corpora, we talk about the common features of Persian in Section 4; and the problems will be discussed in Section 5, as the relevant solutions will be proposed to resolve the problems. Finally, Section 6 summarizes the paper.

2 Features of Developing a Corpus

There are some features that should be taken into account while developing a corpus (McEnery and Wilson, 1996). The first feature is the *orientation* to the language or variety to be sampled in the developed corpus, without having bias. The second one is selecting some *criteria* such as the mood of the text whether the language originates in speech or writing or in electronic form, the type of text whether being a book or a journal. The third feature is the *samples* of the language for the corpus. The corpus might consist of the entire documents, or transcriptions of speech events; but the samples should be limited in size to make it possible for the system to process the data. The fourth feature is that the samples should be a good *representative* for the language. The fifth one is that the corpus should be *balanced* to cover all properties of different kinds of text it contains.
While web pages are likely to be the most immediately accessible source of material, it is possible to handle the massive data sets efficiently to develop corpora. Such materials are problematic when we have a collection of data from various sources.
In this paper, we present the common problems in building Persian corpora from online resources. Before talking about the features of this language and the difficulties to build a corpus for it, we talk about the developed corpora in the next section.

3 Developed Corpora for Persian

There are some corpora for Persian in which they are developed for special purposes. The corpora vary in size and the genres they include, so some of them do not totally cover all the features presented in Section 2. Mostly, the developed corpora for Persian are in plain texts.
FLDB[1] is a linguistic corpus which contains 3 million words in ASCII format released by Assi (1997) at the Institute for Humanities and Cultural Studies. The recent version of the database in 1256 character code page is named PLDB[2] (Assi, 2005) which includes more than 56 million words. This database comprises contemporary literary books, articles, magazines, newspapers, laws and regulations, transcriptions of news, reports, and telephone speeches for lexicography purpose. One advantage of this database is that to each word four linguistic knowledge is attached at once namely phonetic, syntactic, semantic, and lemma tags. The syntactic tag set that has been used is a set of 44 tags (Assi and Abdolhosseini, 2000). No information is available for semantic tags at the moment. Ghayoomi (2004) developed a corpus from 6 months of Hamshahri newspaper on-line archive with more than 6.5 million words. Darrudi (2004) developed another corpus from 4 years of Hamshahri newspaper on-line archive which comprises more than 37 million words. Although the two

[1] Farsi Linguistics DataBase

[2] Persian Linguistics DataBase

corpora are derived from the same source, they do not cover each other as they belong to different dates. Taghiyareh (2003) comprised a 2.5 million words text collection that contains laws and regulations passed by the Iranian Parliament. Bijankhan et al. (2004e) developed a corpus named 'Peykareh' (Text Corpus) in which it includes approximately 38 million words. This corpus consists of newspapers, books, magazines, articles, technical books, and also transcription of dialogs, monologues, and speeches for language modeling purpose. The corpus is going to be expanded up to 100 million words. It should be added over 10 million words of this corpus is tagged semi-automatically (including both automatically and manually) based on EAGLES standard at the moment.

Shiraz corpus is a bilingual parallel tagged corpus which consists of 3000 Persian sentences with the corresponding English sentences. The corpus is collected from Hamshahri newspaper on-line archive. All the sentences are manually translated at CRL[3] of New Mexico State University (Amtrup, 2000).

There are some corpora for Persian speech. One is FARSDAT[4] offered by ELRA[5] and developed at RCISP[6] with the code S0112 as a Persian speech database for phonetic modeling purpose. It consists of 405 sentences read aloud by 304 Persian native speakers chosen from ten different dialectal areas in Iran (Bijankhan et al., 1994; 2004d). LDC[7] has offered two telephone speech corpora in which one is developed by OGI[8] called OGI Multilingual Corpus with the code of LDC94S17 for speech recognition purpose involved 175 calls; and the other one developed by LDC called CALLFRIEND Farsi with the code of LDC96S50 for language identification purpose involving 109 calls. TFARSDAT[9] is the Persian monologue telephone speech database. 64 Persian native speakers have recorded 7:56:7 hours of monologue telephone speech for speech recognition and language identification purposes (Bijankhan et al., 2003; 2004a). The Persian dialog telephone database includes 100 hours of dialogs between 200 Persian native speakers (Bijankhan et al., 2004c). The large Persian speech database includes over 1000 hours of speech recorded by 100 Persian speakers from ten dialects (Bijankhan et al., 2004b). The Persian Telephone Conversation Corpus developed at RCISP is a 100 distant calls of telephone conversation from ten different dialectal areas of Iran for speech recognition and language identification purposes.

Among the corpora mentioned above, some are annotated automatically, manually; and semi-automatically.

4 Properties of Persian

Persian is a member of the Iranian branch of the Indo-European languages which has many features and properties in common with other members in term of morphology, syntax, the sound system, and the lexicon. Even though the scripts of Persian and Arabic are the same,

---

[3] Computing Research Lab

[4] FARsi Spoken language DATabase

[5] European Language Resources Association

[6] Research Center for Intelligent Signal Processing

[7] Linguistic Data Consortium

[8] Oregon Graduate Institute of Science & Technology

[9] Telephone FARsi Spoken language DATabase

Table 1. List of Persian Letters

Name	Letter Form				Phonemes	Example				Unicode
	Non-joiner	Joiner				Non-joiner	Joiner			
		Begin	Middle	End			Begin	Middle	End	
Hamzeh	ء ، أ ، ئ	ئـ	ـئـ	أ ، ئ	ʔ	املاء ، رأس	رئیس	مسئول	هیأت، شیئ	u0621, u0623, u0626
Alef with tild	آ	-	-	ـآ	ʔā	آب	-	-	مأخذ	u0622
Alef	ا	-	-	ـا	ā, a , ʔa, e , ʔe, o, ʔo	دارو ، اکرم، امروز، اروپا	-	-	شما	u0627
Beh	ب	بـ	ـبـ	ـب	b	کتاب	برادر	سبک	سبب	u0628
Peh	پ	پـ	ـپـ	ـپ	p	کپ	پیر	سپاه	لامپ	u067E
Teh	ت	تـ	ـتـ	ـت	t	فوت	توپ	متر	دست	u062A
Seh	ث	ثـ	ـثـ	ـث	s	میراث	ثریا	مثبت	حدیث	u062B
Jim	ج	جـ	ـجـ	ـج	ʤ , ǰ	کاج	جار	لهجه	هویج	u062C
Cheh	چ	چـ	ـچـ	ـچ	ʧ , č	قارچ	چاپ	کچل	گچ	u0686
Heh	ح	حـ	ـحـ	ـح	h	لوح	حرم	صحرا	صلح	u062D
Kheh	خ	خـ	ـخـ	ـخ	x	سوراخ	خاله	صخره	سیخ	u062E
Dāl	د	-	-	ـد	d	در	-	-	سبد	u062F
Zāl	ذ	-	-	ـذ	z	ذلت	-	-	لذت	u0630
Reh	ر	-	-	ـر	r	باران	-	-	سراسر	u0631
Zeh	ز	-	-	ـز	z	روزنامه	-	-	سبز	u0632
Zheh	ژ	-	-	ـژ	ʒ , ž	ژاله	-	-	پژمرده	u0698
Sin	س	سـ	ـسـ	ـس	s	رأس	سبد	پسر	مگس	u0633

Name	Letter Form				Phonemes	Example			Unicode
	Non-joiner	Joiner				Non-joiner	Joiner		
		Begin	Middle	End			Begin	Middle	End	
Shin	ش	شـ	ـشـ	ـش	ʃ , š	موش	شیئ	کشک	کش	u0634
Sād	ص	صـ	ـصـ	ـص	s	خاص	صبح	مصلح	خالص	u0635
Zād	ض	ضـ	ـضـ	ـض	z	مرض	ضایع	مضطرب	مریض	u0636
Tā	ط	طـ	ـطـ	ـط	t	احتیاط	طناب	قطر	رابط	u0637
Zā	ظ	ظـ	ـظـ	ـظ	z	محفوظ	ظالم	منظم	حفظ	u0638
ʻeyn	ع	عـ	ـعـ	ـع	ʔ	اشیاع	علم	معلم	خلع	u0639
Ghein	غ	غـ	ـغـ	ـغ	q	کلاغ	غایب	اشتغال	جیغ	u063A
Feh	ف	فـ	ـفـ	ـف	f	معروف	فکر	کفر	کیف	u0641
Ghāf	ق	قـ	ـقـ	ـق	q	بوق	قاب	مقیم	مشق	u0642
Kāf	ک	کـ	ـکـ	ـک	k	ساک	کمد	عکس	سبک	u06A9
Gāf	گ	گـ	ـگـ	ـگ	g	بزرگ	گنجه	مگس	سگ	u06AF
Lām	ل	لـ	ـلـ	ـل	l	کال	لذیذ	علم	حل	u0644
Mim	م	مـ	ـمـ	ـم	m	چرم	مادربزرگ	کمک	ظالم	u0645
Nun	ن	نـ	ـنـ	ـن	n	باران	نما	منظم	من	u0646
Vav	و	-	-	ـو	v , o, ū , ow, -[10]	تاو، دو، روز، اوج	-	-	تاوا، نوک، بوسه، موز، خواهر	u0648
Heh	ه	هـ	ـهـ	ـه	h , -	ده	هراس	مهم	یه، خانه	u0647
Yeh	ی	یـ	ـیـ	ـی	y , ī	چای	مریم	سیاه	ماهی	u064A

[10] The signe '-' means that it is not pronounced so it has no phonetic representation.

Persian is not comparable in many aspects to Arabic which is from Semitic family. The Persian alphabet is a modified version of the Arabic alphabet. The number of Persian letters is 32, but 28 for Arabic. Persian has four more letters than Arabic: 'پ', 'چ', 'ژ', and 'گ'.
Like Arabic, Persian letters have joiner or non-joiner forms based on the position that these letters appear in a word. The variability of the forms for most of the letters is four. In three positions, the beginning, the middle, the end of a word, the letters are appeared in joined form attached to the neighbor letter(s). The other form is non-joiner which is isolated appears at the very end of a word. For instance the letter '*EYN* can be appeared as 'عـ' for the beginning, 'ـعـ' for the middle, 'ـع' for the end joiner, and 'ع' for the non-joiner of the same letter. Table 1 shows all Persian letters and their various written forms along with their Unicode and an example.
Among the letters, there are some which have only two forms: non-joiner and end-joiner. These letters namely 'ا - آ', 'د', 'ذ', 'ر', 'ز', 'ژ', and 'و' are the ones that can not join to the next letter. For instance the letter *DĀL* is written as 'د' for non-joiner and 'ـد' for the end joiner of a word. It is needed to mention that if either of these letters comes in the middle of a word, the next letter will have the same situation as the beginning of a word.
Persian alphabet is more appropriate to the Arabic sound system and less suitable for Persian. For instance 'ز', 'ذ', 'ض', and 'ظ' are four alphabets both in Persian and Arabic, but all pronounced the same /z/ in Persian and differently in Arabic, i.e. there are different letters for a sound in Persian. Table 2 represents various alphabets pronounced the same in Persian. Although these sets of alphabets are written differently and there is no difference in their pronunciations, they make differentiations in the meanings of words for instance the words ثواب /savāb/ and صواب /savāb/ in which the former means 'reward' and the latter 'right action'.

Table 2. List of Persian letter with same pronunciation

Pronunciation	Alphabet
/t/	ت
	ط
/s/	ث
	س
	ص
/h/	ح
	ه
/z/	ذ
	ز
	ض
	ظ
/q/	غ
	ق

It is also possible to have more than one sound in Persian for a letter; like 'و' in these examples which are underlined: 'نانوا' /nānv̲ā/ 'baker', 'دو' /do̲/ 'two', 'اوج' /o̲wǰ/ 'climax', 'روز' /rūz̲/ 'day', and sometime it is written but not pronounced such as 'خواهر' /xāhar/ 'sister'. So there is a little correspondence between Persian letters and sounds.
Persian writing system for texts is right to left and for numbers left to right, the same as Arabic; but quite contrary to the European languages that have a left to right writing system

both for texts and numbers. The writing system is problematic when within a text there is a number and the sentence should be processed. The system should first process the words right to left and the numbers left to right.
The Persian vocabularies have been greatly influenced by Arabic and to some extent by French while a great amount of words are borrowed from these languages.
Talking about Persian syntax, only verbs are inflected in the language and the number of inflections is six. The subjective mood is widely used. It is a Subject Object Verb (SOV) language with a free word order on the constituent level. This language does not make use of gender; not even the third person of he or she distinctions that exists in English (Assi, 2004; Ghayoomi and Assi 2005). In this language the indicator 'را' /rā/ is used to determine the object.
Persian has three short vowels ' َ ', ' ُ ', ' ِ '; and three long vowels 'ا - آ', 'او', 'ای'. The short vowels are pronounced but mostly not written. The *Ezāfeh* morpheme, the short vowel /e/ used to link the constituents' components together, is also pronounced but not written as we will have more about it in the following subsection. Table 3 shows both short and long Persian vowels including their Unicode and examples.

Table 3. List of Persian vowels

Vowel	Pronunciation	Example	Unicode
َ	a	مَهر	u64E
ُ	o	مُهر	u064F
ِ	e	مِهر	u0650
آ - ا	ā	آمار ، مار	u0622, u0627
او	ū	مور	u0648
ای	ī	میز	u06CC

Persian script, the same as Arabic, has no upper case or lower case letters contrary to the European languages. So, it is not possible to identify proper names or foreign names, and words from other languages easily. Moreover, it is not easy to create acronyms too.
Dot on the above or below of a letter is a distinctive factor for writing or typing Persian letters. For instance the difference between 'ت' and 'ث' is the number of dots, two or three above the letter, which makes their pronunciations different. Besides the number of dots, their position is also important. For example these beginning joiner letters 'نـ', 'تـ', 'ثـ', 'بـ', and 'پـ' pronounce differently. So, the changes in number of dots on letters make the word ambiguous as in words 'بتا' /betā/ 'beta' and 'بنا' /banā/ 'building'. They differ only in the middle letter 'تـ' that has two dots and 'نـ' that has one.
There are some characters that are directly borrowed from Arabic scrip. *Tanvin* is presented to any of these forms: ' ً ' /ʔan/, ' ٌ ' /ʔon/, ' ٍ ' /ʔen/. The other character is diacritic mark ' ّ '. Some borrowed words from Arabic have made Persian to borrow the structure of the words themselves too, such as *Short Alef* ''ی' /ā/ in ''موسی' /mūsā/ 'Moses'; and 'الـ' in 'بالقوه' /belqovve/ 'potential'.
Space is a word boundary between the words in a sentence. Also there is another space named "pseudo-space" as a boundary inside the word. For instance, in the word 'بین‌المللی' /beynolmelali/ 'international' there is a pseudo-space between 'ن' and 'ا'. If the pseudo-space is not available, the letters will be joined to each other.
Since character codes play the most important rolls than the letters itself while typing in computer, it should be said that there are standard codes for Persian characters. In 1993, a

standard 8 bit code for information exchange was approved. In 1995 the keyboard standard layout for Persian was approved; and in 1996 the Unicode standard, the 16 bit code, was approved (Fahim-Niya, 2002).

4.1 *Ezāfeh* Construction

Ezāfeh is an unstressed vowel /e/ considered as a morpheme which is pronounced but not written in Persian. Its function is making links between some constituents building a noun phrase in which this phenomenon is named "*Ezāfeh* construction". Since *Ezāfeh* does not have a graphical representation, text processing would be a tough task specially parsing. What we are sure is that this construction appears in 8 positions as Kaynemuyipour (2000; 2006) determined:

- a noun before another noun: 'کیف چرم' /kif e čarm/ 'leather bag';
- a noun before an adjective or adjectival clause: 'سگ سیاه' /sag e siyāh/ 'black dog', 'عکس چاپ‌شده در روزنامه' /ʔaks e čāpšode dar rūznāme/ 'the picture published in the newspapers';
- a noun before a possessor (noun or pronoun): 'کتاب مریم' /ketāb e maryam/ 'Maria's book', 'the book of Maria';
- an adjective before another adjective in a noun phrase: 'سگ سیاه بزرگ' /sag e siyāh e bozorg/ 'big black dog';
- some prepositions before nouns: 'زیر کیف' /zir e kif/ 'under the bag' but not 'در گنجه' /dar ganǰe/ 'in closet';
- a pronoun before an adjective: 'من پیر' /man e pir/ 'old me';
- first names before last names but sometime it drops: 'مسعود قیومی' /masʔūd e qayyūmi/;
- a combination of above: 'سگ سیاه بزرگ مریم' /sag e siyāh e bozorg e maryam/ 'the big black dog of Maria'

But there is a possibility to have some compounds with post-nominal adjectives in which *Ezāfeh* will not be used (Ghomeshi, 1996) such as 'مادربزرگ' /mādar-bozorg/ 'grandmother'; 'آب پرتغال' /ʔāb-porteqāl/ 'orange juice' and 'آب هویج' /ʔāb haviǰ/ 'carrot juice' but not 'آب سیب' /ʔāb e sib/ 'apple juice'; 'سیب‌زمینی' /sibzamini/ 'potato'; the different words in Persian which means all 'cousin' in English like 'پسرخاله' /persar xāle/. One assumption to dropping *Ezāfeh* could be treating this matter collocational; i.e. such constructions are transforming from compound to collocation.

5 Problems in Corpus Building for Persian

While developing a corpus for Persian from written text, we face several problems based on the specific features of the language mentioned in the previous section. The problems could be solved automatically or manually, but some of them might be left unsolved because it needs a lot of time and energy to be resolved. In this section the problems we faced in our experience while constructing a corpus are listed and the possible solutions to fix these problems are discussed.

5.1 Encoding Issues

The control characters for both Persian and Arabic are very similar to each other but with some differences. One discrepancy is having four more letters in Persian than Arabic; the other one is that the written texts sometimes employ Arabic or ASCII characters as well beside the range of Unicode characters designed for Persian. Hence, the letters 'ک' and 'ی' can be expressed by either the Persian Unicode encoding (u06a9 for 'ک' and u06CC for 'ی') or by the Arabic Unicode (u0643 for 'ك'' and u064A for 'ي') (Megerdoomia, 2004).
Most operating systems have Arabic codes as their default and only codes for 'پ', 'چ', 'ژ', 'ک', 'گ' and 'ی' are added to them to be compatible with Persian. In addition, the texts that are typed on MS Word 97 and FrontPage 97 for web pages use the 1256 control characters; while the MS Word and FrontPage 2000 use the Unicode control characters. As a result, the problem of mixing code pages in typing leads to have the texts with mixed codes.
Since character codes play the most important roll in typing and sorting than the letters themselves, the mixed Arabic and Persian codes make both processing and searching through the text difficult. Fortunately, since the number of these characters is limited, the problem could be solved with a short programming script to convert these letters and have a uniformed text. It should be added that there is a corresponding mapping of Arabic 1256 code to a particular Unicode character based on standard ISIRI[11] 6219 for Persian. So far the problem of typing solved but not sorting. The sorting of Persian letters is as follows:
الف، ب، پ، ت، ث، ج، چ، ح، خ، د، ذ، ر، ز، ژ، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ک، گ، ل، م، ن، و، هـ، ی.
Using the sorting of Arabic letters Persian will mess up the order of Persian letters as the position of 'و' and 'هـ' is swapped in Arabic alphabet and the four letters namely 'پ', 'چ', 'ژ', and 'گ' do not exist in Arabic. So that the following incorrect order will be the result:
الف، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ک، ل، م، ن، هـ، و، ی، پ، ژ، گ، چ.

5.2 Internal Word Boundaries

One of the biggest issues in processing Persian texts is the internal word boundary that should be presented with a zero-width non-joiner space named pseudo-space. While typing a word, typists do not care about pseudo-space and they enter a white space instead or completely ignore the internal word boundary. When white space is used instead of the pseudo-space such words are treated as separate words which causes problem in text processing. As an example 'بین‌المللی' is typed 'بین المللی' which is considered as two separate words 'بین' and 'المللی'. Such a problem has a negative effect on the frequency distribution of words in text processing. The similar problem also happens when the pseudo-space is ignored, because the frequency of such words will be distributed between different typing styles.
Moreover, optionality of the internal word boundary raises problems in the analysis of detached morphemes such as 'ها -' /-hā/ (a plural morpheme), 'ای -' /-i/ (post-determiner), 'می -' /mi-/ (a present/past continuous morpheme), 'تر -' /-tar/ (a comparative suffix), or 'ترین -' /-tarin/ (a superlative suffix). The inflectional morphemes can appear either as bound

[11] Institute of Standards and Industrial Research of Iran

and joined to the host as free affixes separated by a pseudo-space, or separated with an intervening white space as presented in Table 4. Choosing any of these styles depends on the typist's typing style.

Table 4. Different types of word boundaries for affixes

Affix	White space	Pseudo-space	Attached
ـ ها	پول ها	پول‌ها	پولها
ـ ای	خانه ای	خانه‌ای	خانهای
می ـ	می گوید	می‌گوید	میگوید
ـ تر	بزرگ تر	بزرگ‌تر	بزرگتر
ـ ترین	بزرگ ترین	بزرگ‌ترین	بزرگترین

To solve this problem, we decided to change all these three forms to pseudo-space form which is more similar to the original standard format of Persian writing. Since there are a limited number of affixes in Persian, this solution is not a big problem and could be done automatically. Apart from affixes, however, there are more samples that have this problem, specially in the words which inflection and derivation have happened. Using pseudo-space automatically for these kinds of words is not easy as there is a possibility to substitute a psude-space which is not appropriate. Table 5 shows some samples of Persian words that are usually written with or without white space while they should be written with pseudo-space to be considered a single word.

Table 5. Different kinds of word boundaries for derived and inflected words

White space	Pseudo-space	Attached
بین المللی	بین‌المللی	
زبان شناسی	زبان‌شناسی	زبانشناسی
کتاب سرا	کتاب‌سرا	کتابسرا
دانش آموز	دانش‌آموز	
علاقه مند	علاقه‌مند	علاقمند

The white space problem is remained in the compound words such as 'ماشین لباسشویی' /māšin/ /e/ /lebāssšuyi/ 'washing machine'. In tokenization process a compound word could be counted as unseparated words. The compound words are mostly joined together with the *Ezāfeh* morpheme that has no visual representation. Any morphological analyzers for Persian should be able to recognize the various writing styles of words and the *Ezāfeh* morpheme.

Persian has two kinds of prepositions: simple and complex. The complex preposition is composed of the combination of two simple prepositions. Based the study of Abolhasani and Ghayoomi (2006), incorporation happens between the prepositions. They showed different examples to prove the superiority of using pseudo-space in complex prepositions to have a correct processing and analyses of the text in natural language processing applications. We have also benefited of this approach for complex words.

Another word boundary problem of Persian related to complex words. Complex words refer to multi-element forms which consist of separate lexical items that are attached to each other. These attached morphemes such as the preposition 'به' /be/ 'to', the prefix 'هم' /ham/ 'mate', the determiner 'این' /ʔin/ 'this', the postposition 'را' /rā/, or relativizer 'که' /ke/ 'that', may appear attached to the adjacent word or separated by pseudo-space or white

space (Megerdoomian, 2004). Table 6 shows different writing format of the items just named when used with other words.

Table 6. Different kinds of word boundaries for complex words

Word	Type	White space	Pseudo-space	Attached
به	Preposition	به شیوه	به‌شیوه	بشیوه
هم	Prefix	هم کلاس	هم‌کلاس	همکلاس
این	Determiner	این کار	این‌کار	اینکار
آن	Determiner	آن قدر	آن‌قدر	آنقدر
را	Postposition	شرایط را	شرایطرا	شرایطرا
که	Relativizer	چنان که	چنان‌که	چنانکه

Non-joiner letters are another source of word boundary problem in Persian. As it is already mentioned in section 4, some Persian letters including 'ا - آ', 'د', 'ذ', 'ر', 'ز', 'ژ', and 'و' do not join to the next letter at all. Such letters make problems when they appear at the end of a word. Since the forms of the letters do not change, then the typists usually do not care to enter a white space after these letters. For example 'اورا' /ʔūrā/ 'he or she (in accusative case)' is spelled as a single unit; while a white space has to be entered between them to consider them as two separate words, i.e. 'او را'. As shown, 'و' and 'ر' remained unchanged with or without the white space, since 'و' does not join to the next word.

It is hard to find an automatic solution to this problem. One alternative solution is searching manually for these words and separating them as two words. However, this task needs a lot of efforts and also there is no guaranty to find all of them.

The opposite situation can also happen for the words that should be written as a single word, but since they constructed from other words, they might have been written separately. For example the word 'درغیراینصورت' /dar qeyr e ʔin sūrat/ 'otherwise' which is an adverb and should be considered as a single word, in the lexicon the elements that have built this word are available as free morphemes which are 'در' /dar/ 'in', 'غیر' /qeyr/ 'other', 'این' /ʔin/ 'this', and 'صورت' /sūrat/ 'form'. Since we have these free morphemes, problem occurs when space is used between the morphemes which leads to have different writings, i.e., the single word is available in several formats in a corpus: 'در غیراینصورت', 'درغیر اینصورت', 'درغیراین‌صورت'.

Numbers are another big issue in word boundary which have three important features when they appear in a text: one is their left-to-right writing style in digits, their right-to-left writing style in letters, their disjoining to the next letters in digits. One problem occurs when numbers are written in letters such that a number in digits counted as one word can be considered as separate words in letters in case the number is bigger than twenty; such as 123,000 that is written as 'صدوبیست‌وسه‌هزار' (with no white space), 'صد و بیست و سه هزار' (with white space), or 'صدوبیست‌وسه هزار' (with white and pseudo space).

The third feature is problematic when no white space is used between a number and the next word, such as '۱۲۳بشکه' /sad o bist o se boške/ '123 barrels'.

5.3 Writing Style

There exits three kinds of language varieties: standard, super-standard, and sub-standard. The standard language both for oral and written is the one as the official language used in press, mass media, formal communications and correspondences, etc. The super-standard language which has more cultural and artistic values is used in literary texts and scientific text books. The sub-standard language is used in SMSs, blogs, and also used as a colloquial and casual speech in daily conversation in the society. Slang and taboo are sub-branches of this language variety.

We can not always have a rigid and clean border between these varieties as there is a possibility to have an amalgamation of them within a text. For example, in a storybooks or novels, there is a possibility that the writer shifts from one style to another.

Such a language variability makes the text processing difficult. As an example, the word if which is written 'اگر' /ʔagar/ in standard and super-standard texts is changed to 'اگه' /ʔage/ in colloquial form as oral data or 'گر' /gar/ in literature and poems.

5.4 Linguistic Creativity

Aside from the usual language changes and creativities over time which are the properties of live languages, recent communication technologies such as SMS (Short Message Service) and Internet have speeded them up. Entering a text to a mobile phone is done through 9 keys such that each key carries at least three characters and typing a complete word, especially Persian compound words, would be a very difficult task and time consuming for users; as the result pressing less keys would be more convenient for them. This is the condition that users create words from the original word which has less number of characters and takes less time to enter consequently. For example, 'زنگیدن' /zangidan/ is one of the created words in SMS texts which used instead of the compound verb 'زنگ‌زدن' /zang zadan/ 'to call'. It should be added that this linguistic phenomenon is not limited in SMS texts as it has entered to the chat-rooms and blogs too. Surely the processing of texts out of these sources beside other sources would be a real big challenge.

The data out of chat-rooms and blogs are linguistically valuable but processing of them is a tough task as there is a free writing style and a lot of misspelled words and grammatical mistakes could be found. Besides the problems, there would be more interesting linguistic data when some facial icons are used in texts to express emotions as they have meaning semiotically.

Even some writers of books, mostly literary and novels belonging to super-language class, use their own personal style and they will have creativities in the accepted orthography of the words with no changes on the pronunciation, meaning, or syntactic function. For example the words 'حتی' or 'حتّی' /hattā/ 'even' and 'حتماً' /hatman/ 'certainly' which have the accepted orthographies are written as 'حتا' and 'حتمن'.

5.5 Homographs and Homonyms

Like any other languages, Persian also contains different ambiguities in the lexicon. However, because of two important features of Persian letters, the numbers of homographs

and homonyms surprisingly increase.
One of these features is writing no short vowels which causes to have a lot of ambiguities and homographs in the language. Any morphological analyzers for Persian should be able to recognize and disambiguate such words. For example the homograph 'کند' could be pronounced any of these along with their POS taggings: /kand/ 'picked' (Verb, Past Tense) and /kanad/ 'picking up' (Verb, Present Continuous Tense), /konad/ 'doing' (Verb, Present Tense), /kond/ 'slow' (Adv), and /kond/ 'blunt' (Adj). On the other hand, it makes problems in tokenization process to extract the frequency, because these five words would be counted as one, and the result is an unreliable statistics for the frequency of such words. This is also one of the problems which are very difficult to solve. Enriching words with some linguistics knowledge sush as POS tags could resolve most of these problems.

Another problematic feature is the proper names that can not be distinguished with capital litters, since capitalization does not exist in Persian. For example the word 'آذر' /ʔāzar/ is the name of the 9th month in Persian calendar; a girl name, and it also means fire. These words could be disambiguated based on the local context they are used in. This feature also makes the Persian name entity recognition more difficult.

5.6 Borrowed Characters from Arabic

The borrowed characters from Arabic make the processing of a corpus more challenging, because most of the typists do not care about these characters at all in typing. There are four different Arabic characters which cause the most challenging part of corpus development for Persian: *Tanvin*, diacritic mark, *Hamzeh*, and *Short Alef*.
Some typist use *Tanvin* and some do not care. This different typing style causes to have two separated words in the vocabulary for a single word like 'فعلاً' /feʔlan/ 'yet' and 'فعلا'. This problem becomes more critical when using *Tanvin* distinguishes some words from the other. For example a word which is typed 'جدا' could be 'جدا' /ǰodā/ 'separate' or 'جداً' /ǰeddan/ 'really'.
The other character is the diacritic mark 'ّ' which can be written or ignored such as 'معلّم' and 'معلم' /moʔallem/ 'teacher'. Like *Tanvin*, the diacritic mark can make a distinction between words' meaning. For example the word that is typed 'بنا' could be either 'بنا' /banā/ 'building, base' or 'بنّا' /bannā/ 'bricklayer'.
Hamzeh is another Arabic character which optionally appears in some words. This character can be used at the end of some words such as the word 'املا' /ʔemla/ 'dictation', which can be written as 'املاء'. This character is used in the middle or end of words having any of these long vowels as base: 'ا', 'ی', and 'و'. Having *Hamzeh* or not in Persian is a discussion among linguists. The following words are three examples based on the first group's assumption who does not believe in having *Hamzeh* in which they use the corresponding long vowel instead such as the words 'مساله' /masale/ 'problem', 'رییس' /reyis/ 'boss', and 'مومن' /momen/ 'believer'; while these words are written as 'مسئله' /masʔale/, 'رئیس' /reʔis/, and 'مؤمن' /moʔmen/ basd on the assumption of the second group who believes in having *Hamzeh* in Persian.
Having 'ی' or *Hamzeh* is not the end of story because in some words both 'ی' and *Hamzeh* can be removed. As an example the word 'mirror' could be written in three different forms:

with 'ی' as 'آیینه' /ʔāyine/, with *Hamzeh* as 'آئینه' /ʔāʔine/, and without the 'ی' or *Hamzeh* as 'آینه' /ʔāyene/.
Like other borrowed characters, some typists use *Short Alef* while they are writing a text and some do not care. This has made a problem in word frequency distribution. For example 'Moses' could be written as either 'موسیٰ' or 'موسی'.
The borrowed characters make problems in the tokenization process of a corpus for Persian. For solving this problem we removed all three types of *Tanvin*, the diacritic mark, *Hamzeh*, and also *Short Alef* if they appeared in the text.

5.7 Various Orthographies for Words

Persian orthography is not completely ruled based and standard. Even the "Academy of Persian Language and Literature" recently has published a book for this purpose. But after giving the rules, a list of exceptions are presented; so it doesn't make a lot of help to remove problems in processing a corpus. Different typing styles of typists have added problems too.
One of these problems is caused by the nature of 'ا - آ'. As presented in Table 1, the letter 'آ' appears quite often at the very beginning of words and rarely in the middle of the words. But the letter 'ا' appears at the beginning, middle or end. However, since these two letters are pronounced the same, some typists do not care about and only use 'ا'. For instance the word 'relax' that should be written as 'آرام' /ʔārām/ is written as 'ارام'.
The variability of using *Hamzeh* or 'ی' in some words also causes another problem in corpus as we saw in the previous subsection.
It is also possible to have both problems in one single word which make the text processing more difficult that before. For example a word meaning 'American' is spelled in four different forms: 'امریکائی' /ʔemrikāʔi/, 'امریکایی' /ʔemrikāyi/, 'آمریکائی' /ʔāmrikāʔi/, and 'آمریکایی' /ʔāmrikāyi/ and they are counted as four different words.
To overcome such inconsistencies of Persian, Ghayoomi (2004) used the highest frequency of typings as a guide to make decision about the default spelling, and he has manually replaced the various spellings to the selected spelling.

In Persian, *Ezāfeh* is represented in two different forms based on syllabic restrictions existing in Persian[12]. One form is its appearing after a consonant as /e/ which is not mostly written like 'کیف مریم' /kif e maryam/. The other form is when *Ezāfeh* appears after a vowel in which the intermediary morpheme 'ی' which is a consonant will be required. The problem occurs when there are four types of writing for *Ezāfeh* in a corpus. One is appearing the consonant 'ی' with a pseudo or white space at the end of a word: 'خانهی' or 'خانه ی' /xāne y e/ 'the house of'. Here the intermediary morpheme is inserted to follow the

[12] Persian has three kinds of syllabic restrictions for the sequence of consonants (C) and vowels (V) which are: CV, CVC, and CVCC. If two morphemes are going to glue to each other to build another morpheme, in case the first morpheme is ended to a vowel and the next one starts with a vowel too, then a consonant is required to follow Persian syllabic restrictions. This consonant which is considered as a morpheme has different pronunciations based on the context it appears.

syllabic restrictions. The other typing is appearing a combination of both the intermediary morpheme and the vowel /e/ above the letter 'هـ' such as 'ۀ' in the word 'خانۀ'. Since the code page of 'ۀ' for Persian font is the same as 'ـة' for Arabic, we will have different surface appearances in case a Persian font is used; otherwise 'ۀ' would appear as 'ـة'. The other possibility is not writing *Ezāfeh* neither of these forms 'ی' or 'ۀ'. Having these various forms makes searching through corpora a tough task. To overcome such difficulty we could automatize this problem to some extent.

5.8 Foreign Words

Since Persian alphabet is different from Latin, writing foreign words in Persian is very difficult. There is no standard for it and based on the way the author/writer/typist pronounce the Latin words, they write it based on Persian syllables. Having this phenomenon, for example we could find four kinds of spellings for the word 'intermediate' as 'اینترمدیت', 'اینترمدییت', 'اینترمیدیت', and 'اینترمیدییت'.

6 Conclusion

In this paper we mentioned some of the common problems we experimentally faced in developing a corpus for the Persian language from written text and described our rough solutions to fix the problems.

The source of problems could be the Persian script mixed with Arabic script; Persian orthography; the typing style of typists; the control characters' code pages in the operating systems and word processors; having various linguistic style and creativity in the language. Generally speaking, before processing the Persian corpus, it is needed to preprocess the raw data automatically, manually, and a combination of both by spending energy and time.

7 References

Amtrup, J. W. et al. (2000) "Persian-English machine translation: An overview of the Shiraz project" NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-319)

Abolhasani, Z. and M. Ghayoomi (2006) "Incorporation: Word production of Persian prepositions and its application in computational linguistics". In Proceedings of the 2nd Workshop on the Persian Language and Computer, Tehran University, Tehran, Iran, . pp. 16-24.

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

Assi, S.M. (2004) "Persian language and IT" In Proceedings of the 2nd Workshop on Information Technology and Its Disciplines, pp.85-94, Kish Island, Iran.

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

Assi, M. and M. Hajiabdolhosseini (2000) “Grammatical tagging of a Persian corpus”. International Journal of Corpus Linguistics 5(1), 69-81.

Bijankhan, M. et al. (1994) “FARSDAT: Farsi spoken language database”. In Proceedings of International Conference on Speech Sciences and Technology, Vol. 2: 826-829, Perth, Australia.

Bijankhan, M. et al. (2003) “TFARSDAT: Telephone Farsi spoken language database. EuroSpeech, Geneva: (3): 1525-1528.

Bijankhan, M. et al. (2004a) “Persian monologue telephone speech database: TFARSDAT”. In Proceedings of the 1st Workshop on Persian Language and Computer, Tehran University, Tehran, Iran, pp. 147-148.

Bijankhan, M. et al. (2004b) “The large Persian speech database”. In Proceedings of the 1st Workshop on Persian Language and Computer, Tehran University, Tehran, Iran, pp. 149-150.

Bijankhan, M. et al. (2004c) “The Persian dialogue telephone database”. In Proceedings of the 1st Workshop on Persian Language and Computer, Tehran University, Tehran, Iran, pp. 151-152.

Bijankhan, M. et al. (2004d) “The Persian Speech Database: FARSDAT”. In Proceedings of the 1st Workshop on Persian Language and Computer, Tehran University, Tehran, Iran, pp. 145-146.

Bijankhan, M. et al. (2004e) “The Persian text corpus”. In Proceedings of the 1st Workshop on Persian Language and Computer, Tehran University, Tehran, Iran, pp. 143-144.

Darrudi, E. et al. (2004) “Assessment of a modern Farsi corpus” In Proceedings of the 2nd Workshop on Information Technology and Its Disciplines, pp.73-77, Kish Island, Iran.

Fahim-Niya, F. (2002) Problems in Teaching and Learning Persian Script to Primary Students at the Second Grade. MA thesis. Institute for Humanities and Cultural Studies, Tehran, Iran.

Ghayoomi, M. (2004) Word Prediction in Computational Processing of the Persian Language. MA thesis. Islamic Azad University, Tehran Central Branch, Iran.

Ghayoomi, M. and S.M. Assi (2005) “Word prediction in a running text: A statistical language modeling for the Persian language” In Proceeding of the Australasian Language Technology Workshop, University of Sydney, Australia.

Ghomeshi, J. (1996) Projection and Inflection: A Study of Persian Phrase Structure. Ph.D. dissertation, University of Toronto.

Kahnemuyipour, A. (2000) “Persian Ezafeh construction revisited: Evidence for modifier phrase” In J.T. Jensen and G. van Herk (eds.) Cahiers Linguistique d'Ottawa, Proceedings of the 2000 Annual Conference of the Canadian Linguistic Association, pp. 173-185.

Kahnemuyipour, A. (2006) “Persian Ezafeh construction: Case, agreement or something else” In Proceedings of the 2nd Workshop on Persian Language and Computer, Tehran University, Tehran, Iran.

McEnery, T. and A. Wilson (1996) Corpus Linguistics. Edinburgh University Press.

Megerdoomia, K. (2004) "Developing a Persian part of speech tagger". In Proceedings of the 1st Workshop on Persian Language and Computer, pp. 99-105, Tehran University, Tehran, Iran.

Taghiyareh, F. et al. (2003) "Compression of Persian text for web-based applications, without explicit decompression" WSEAS Transactions on Computers, Issue 4, Vol. 2, pp. 961-966.