Lang Resources & Evaluation (2011) 45:143–164
DOI 10.1007/s10579-010-9132-x

ORIGINAL PAPER

Lessons from building a Persian written corpus: Peykare

Mahmood Bijankhan · Javad Sheykhzadegan · Mohammad Bahrani · Masood Ghayoomi

Published online: 3 November 2010

Abstract This paper addresses some of the issues learned during the course of building a written language resource, called 'Peykare', for the contemporary Persian. After defining five linguistic varieties and 24 different registers based on these linguistic varieties, we collected texts for Peykare to do a linguistic analysis, including cross-register differences. For tokenization of Persian, we propose a descriptive generalization to normalize orthographic variations existing in texts. To annotate Peykare, we use EAGLES guidelines which result to have a hierarchy in the part-of-speech tags. To this aim, we apply a semi-automatic approach for the annotation methodology. In the paper, we also give a special attention to the Ezafe construction and homographs which are important in Persian text analyses.

Keywords Contemporary Persian · Corpus · EAGLES-based tagset · Ezafe construction · Homographs

M. Bijankhan (✉)
Department of Linguistics, The University of Tehran, Tehran, Iran
e-mail: mbjkhan@ut.ac.ir

J. Sheykhzadegan
Research Center for Intelligent Signal Processing, Tehran, Iran
e-mail: sheikhzadegan@rcisp.ac.ir

M. Bahrani
Computer Engineering Department, Sharif University of Technology, Tehran, Iran
e-mail: bahrani@ce.sharif.edu

M. Ghayoomi
German Grammar Group, Freie Universität Berlin, Berlin, Germany
e-mail: masood.ghayoomi@fu-berlin.de

1 Introduction

Persian belongs to the Indo-European (IE) language family and to the Arabic script-based languages in terms of orthography and text processing. As a subclass of the western Iranian Languages, Persian is currently spoken in Iran, Afghanistan, and Tajikistan where it is called 'Persian/Farsi', 'Dari', and 'Tajik' respectively. Persian has been in interaction with other non-Iranian languages such as Arabic, Turkish, Hindi, Mongolian, Latin, Greek, Russian, English, and French. Hussain and Gul (2005) compared status of standardization and development of basic localization applications among 14 Asian languages, and they found Persian to be in a satisfactory status.

To study linguistic phenomena in Persian, we collected various texts in the electronic format and built a corpus as a language resource. The end goal of this article is addressing the lessons learned from building this corpus. As an introduction to the discussion, it is useful to mention some of the available general corpora for Persian. The FARsi Spoken language DATabase (FARSDAT) is one of the sources offered by the European Language Resources Association (ELRA). This corpus consists of the speech of 300 native speakers of Persian chosen from ten different dialectal areas in Iran which is produced by the Research Center for Intelligent Signal Processing (RCISP; http://www.rcisp.com; Bijankhan et al. 1994). The Linguistic Data Consortium (LDC) has designed and collected another linguistic resource for Persian in the OGI Multilingual Corpus which has covered the data of 100 Persian speakers (Muthusamy et al. 1992). This consortium has also produced another telephone database called CALLFRIEND Farsi with the code LDC96S50 in which the telephone conversations of more than 100 Persian speakers are recorded and verified.

The Institute for Humanities and Cultural Studies (IHCS) has produced an online database called the Persian Linguistic DataBase (PLDB; http//www.pldb.ihcs.ac.ir), in which millions of words from the contemporary modern Persian with different genres have been collected such that the syntactic categories of a small amount of data is labeled by means of a set of 44 tags (Assi and Abdolhosseini 2000) and it is manually lemmatized. Moreover, the phonetic tags of the sample texts are assigned semi-automatically. The database includes texts of both formal and informal styles. The Hamshahri Corpus (http://www.ece.ut.ac.ir/dbrg/Hamshahri/) consists of 345 MBs of news texts from the Hamshahri newspaper during the years 1996–2002 developed by the DataBase Research Group (DBRG) of the Faculty of Engineering at the University of Tehran. This corpus is prepared for different information retrieval research areas such that 65 queries and their relevance judgments for the top 100 retrieved documents according to TREC standard were created.

RCISP, in addition to FARSDAT, has produced three spoken corpora and a written corpus. The Telephone FARsi Spoken language DATabase (TFARSDAT) consists of 7 h of read and spontaneous speech produced as monologue by 60 native speakers of Persian from ten different dialectal areas of Iran, segmented and labeled into phonemic, phonetic, and word levels (Bijankhan et al. 2003). The Large FARSDAT is a Persian speech corpus which consists of read aloud speech from the newspaper texts in which 100 Persian speakers have produced, in average, 25 pages

Table 1 The general corpora of the contemporary Persian

Name	Designer	Data type	Size	Function
FARSDAT	RCISP	Microphone speech	25 h	Phonetic modeling
Large FARSDAT	RCISP	Microphone speech	45 h	Speech and speaker recognition
TFARSDAT	RCISP	Telephone read speech and conversation	About 11 h	Speech recognition and caller identification
CALLFRIEND Farsi	LDC	Telephone speech	109 calls	Language identification
OGI multilingual corpus	OGI	Telephone speech	175 calls	Speech recognition
The Persian telephone conversation corpus	RCISP	Telephone conversation	About 37 h	Speech recognition and language identification
Peykare	RCISP	Text	More than 100 million words	Language modeling
PLDB	IHCS	Text	Not reported	Lexicography
Hamshahri corpus	DBRG	Text	345 MBs	Information retrieval

each, and whose speech was recorded by three kinds of microphones. This corpus consists of 45 h of microphone speech (Bijankhan et al. 2004). The Persian Telephone Conversation Corpus includes 100 long-distance calls from ten different dialectal areas of Iran in which each call is about 20 min long and each word is manually labeled phonemically, phonetically and orthographically. In all of these calls, the variety of the subject matter of the conversation is considered and the number of male speakers is twice as many as the female speakers (Sheykhzadegan and Bijankhan 2006). Peykare is a written corpus which contains approximately 110 million words of both written and spoken texts of the Contemporary Persian (CP). This corpus is categorized according to the criteria such as factuality, format, style, and linguistic material. About ten million word tokens of this corpus were selected randomly and labeled according to the EAGLES guidelines. Table 1 illustrates a summary of the available general corpora for Persian along with their function.

This paper is organized in eight sections. After this brief introduction of Persian and the available general corpora, we will talk about the non-/linguistic parameters taken into consideration for sampling frame in Sect. 2. Then in Sect. 3, we will describe how the texts are collected to construct Peykare. In Sect. 4, the tokenization process will be elaborated by defining two linguistic units: multi-unit tokens and multi-token units. In Sect. 5, we will mainly discuss the annotation of Peykare with the help of EAGLES guidelines. In Sects. 6 and 7, two important issues of Persian that should be considered in corpus development are taken into consideration; namely the Ezafe construction and homographs. The paper ends with a summary in Sect. 8.

2 Linguistic preliminaries

In this article CP is in focus. CP is the last era of the modern Persian which has been the formal language of Iran for the last 162 years. To clarify the time intervals in

which CP texts have been published, we have considered the political milestones as a distinctive border. This is because of the strong effects of the political events in each period (as listed below) on the lexical items of both written and spoken Persian, used by the media and the speakers of this language:

- 1847–1906: AD before the period of 'Mashroutiyat' (Constitutionality);
- 1906–1925: from Constitutionality until the first king of the Pahlavi dynasty;
- 1925–1941: from the first king of the Pahlavi dynasty to the second king;
- 1941–1978: from the second king of the Pahlavi dynasty to the Islamic revolution;
- 1978–1988: from the Islamic revolution to the end of the war with Iraq;
- 1988–2006: from the end of the war until 2006 when designing Peykare ended;
- 2006–Present: from 2006 until now when text collecting for Peykare resumed.

In addition to the above metalinguistic parameters, linguistic varieties are also considered as a socio-linguistic parameter. Since 'standardness' (Douglas 2003) and 'formalness' (Hodge 1957) are two complex and fuzzy parameters, we believe that three linguistic varieties can be identified for Persian during the last century: Standard (S), Super-Standard (SupS), and Sub-Standard (SubS) in which for each of them, potentially, there exist Formal (F) and InFormal (InF) styles; thus, twelve varieties of CP can be taken into account. Douglas (2003) has explained the complexities of how to define language varieties based on these parameters in gathering the Scottish corpus. Almost the same situation holds for CP, too. Since the written mode has only been considered in Peykare, and there is no formal style for the sub-standard variety, it can be expected that the texts can be collected for five varieties of CP; namely standard-formal, standard-informal, super-standard-formal, super-standard-informal, and sub-standard-informal.

3 Composition of Peykare

Peykare is a core synchronic general corpus which includes texts of the five above-mentioned linguistic varieties. To have the qualification of being the representative of the language, the corpus is designed in such a way to be comprehensive enough and include different registers so that the 'random error' and the 'bias error' decrease to the minimal level (Leech 2002; Biber 1992, 1993). When the number and the length of the text samples are not enough to estimate the linguistic parameters, random error increases; and when the text samples do not cover a wide domain of registers, the bias error increases. The first requirement, nowadays, is easy to achieve because of the vast amount of sample texts freely available on the Internet; but satisfaction of the second requirement is not easy for CP. However, the real challenge is how well a corpus represents the register diversity given that there are marked linguistic differences across registers (Biber 1993).

Peykare consists of 35,058 text files, each of which includes either a full text or a random sample of a full text. The size of each text is a chain of at least one thousand words; while the size of some newspaper texts which include short news or commercials is less. Two kinds of criteria have played the main function in the

process of choosing data for Peykare: a linguistic criterion to distinguish the five linguistic varieties of CP; and the non-linguistic criteria which include the variables depicting the communicative function of the texts among the language users such as time, mode, factuality, and the medium (Al-Sulaiti and Atwell 2006; Kučera 2002; http://www.corpora.fi.muni.cz/ske/doc/urg.html). Peykare, for the time being, includes the texts produced during the years 1978–2003. Of course, the works of some famous writers that are out of this time span, such as Hedāyat, are included as well.

Mode shows the way the linguistic data is conveyed. Peykare merely includes the written texts. Spoken corpora have been planned and produced in projects such as FARSDAT, TFARSDAT, and the Large FARSDAT. Generally, written texts can be classified into two groups (Atkins et al. 1992): written to be read (WR), and written to be spoken (WS). The Statistical Center of Iran (http://www.sci.org.ir/) has reported that the literacy rate of 15–24 year old people has increased from 84.6% in 1990 to 95.4% in 1998. Furthermore, because of the significant increase of educational facilities, technical and occupational services, the number of readers of books, magazines, and newspapers has increased as well. As a result, the size of WR texts in Peykare is greater: 87% of WR versus 13% of WS on the whole.

Factuality is a variable with the values of fiction and non-fiction (Kralik and Šulc 2005). Since Persian literature is replete with fiction texts in prose and poem, and they have a tremendous effect on the written and spoken Persian, it covers a significant proportion of Peykare. This is against Sinclair's (1987; 1–8) opinion since he believes that the proportion of literary works should be low in a general corpus; but it is a fact that Persian as a communicative language and the language of science is full of literary clichés. On the whole, about 22.88% of the texts are fiction, 40.20% non-fiction, and about 36.8% a combination of both fiction and non-fiction. A considerable amount of fiction and non-fiction texts which are read by many readers are translated from foreign languages, mostly from English and Arabic. The vocabulary and the syntactic structures of translated texts are completely marked.

Medium is a variable which shows in what format the contents of the texts are published. The medium in Peykare are: books, magazines, periodicals, newspapers, web pages, compact disks, unpublished texts, and manuscripts.

The content of Peykare is categorized under 24 different registers as represented in Table 2 along with the linguistic criteria used for text collection. We have classified these 24 registers under the five linguistic varieties and evaluated the cross-register differences for the five levels. To test the extent of cross-register differences for language variety, twelve linguistic parameters were chosen namely: first and third personal pronouns; the most frequent nouns, preterite verbs, and indicative present verbs; prepositions; verb-locative construction; question words; relative, complement, and conditional subordinate clauses; passive construction I and II. Each parameter consists of a finite set of words obtained by searching through the tagged Training Corpus (TC) using the Searchdata tool which looks for morphological and syntactic structures via regular expressions. The most frequent words were obtained by sorting words with their tags in descending order. Relative and complement clauses were obtained simply by the following orthographic regular expression (X is a string of allographs):

Table 2 Text distribution of Peykare in terms of register, language variety, mode, factuality, and medium

Register	Language variety	Mode	Factuality	Medium	Sample size (word)
Novel	SupS, InF	WR	Fiction	Book	11,009,000
Dramas (1)	S, F	WS	Fiction	Book	6,877,000
Academic	SupS, F	WR	Non-Fiction	Book	12,621,000
Newspaper	S/SupS/SubS, F/InF	WR	Non-/Fiction	Newspaper	26,440,000
Popular, general	S, F	WR	Non-/Fiction	Magazine	10,420,000
Periodicals	SupS, F	WR	Non-Fiction	Book	12,380,000
Education	S, F	WR	Non-Fiction	Book	2,549,000
Manuals	S, F	WR	Non-Fiction	Book	930,000
Constitutions	S, F	WR	Non-Fiction	CD	425,000
Government projects	S, F	WR	Non-Fiction	Web page	335,000
Regulations	S, F	WR	Non-Fiction	CD	1,140,000
Travel Report	S, InF	WR	Non-Fiction	Personal/web page	525,000
Year book	S, F	WR	Non-Fiction	Book	417,000
Correspondences	S, F	WR	Non-Fiction	Archive	84,000
Minutes	S, F	WR	Non-Fiction	Archive	23,000
Daily notes	S, InF	WR	Non-Fiction	Personal	85,000
Presonal letters	S, InF	WR	Non-Fiction	Personal	65,000
Brochures	S, F	WR	Non-Fiction	Web page	1,140,000
Lectures, non-prepared	S, InF	WR	Non-Fiction	Radio/TV	1,600,000
Lectures, prepared	S, F	WS	Non-Fiction	Radio/TV	200,000
Drama (2)	S, InF	WS	Fiction	Radio	5,000,000
Broadcasting news	S, F/InF	WS	Non-Fiction	Radio/TV	785,000
Conversation	SubS, InF	WS	Non-Fiction	Telephone	735,000
Interviews	S, InF	WS	Non-Fiction	Radio/TV	2,900,000

که (را) {ئی|ای|یی} X

Passive construction I comprises a past participle followed by the inflected forms of the passive auxiliary-verb ‘شدن’ /šodan/ ‘to become’. Passive construction II comprises any other words followed by the passive auxiliary verbs.

To examine cross-register differences in the distribution of the twelve parameters, the text files of each register in Peykare were divided into subtexts with 4,000 words length. 100 subtexts were chosen randomly for 15 registers whose total size is larger than 400,000 words. The number of subtexts for registers of governmental projects, correspondences, minutes, personal letters, and prepared lectures were 83, 21, 5, 15 and 44, respectively. Frequency of the words pertaining to each linguistic parameter was counted for all subtexts; and then the ANOVA test was used to differentiate the levels of varieties with mean, standard deviation, and *F*-value. The conclusion, as follows, suggests that there exist significant differences among the levels of varieties:

- Conditional subordinators, first person pronouns, prepositions, present indicative verbs, relative clauses and passive construction I are the best discriminators of the sub-standard informal against the other kinds of varieties.
- Irrespective of standardness, first and third person pronouns, preterite verb, verb-locative construction, and WH questions are the best discriminators for the formal and informal styles.
- Nouns and passive construction I are the only linguistic parameters discriminating all varieties against each other.
- Irrespective of formality, prepositions, indicative present verbs and passive constructions II are the best discriminators for standard and super-standard varieties.

4 Tokenization

Ghayoomi and Momtazi (2009) and Ghayoomi et al. (2010) have described the problems to deal with in developing a corpus for Persian including the tokenization problem. Word boundary is the most challenging issue for tokenization. In Persian texts, a word can be considered as a chain of letters which makes up at least one free morpheme. Typists intuitively recognize words according to this definition, like other literates; while typing texts, however, they do not separate words which results in orthographic variations (Buckwalter 2005).

At least two reasons can be considered for the orthographic variations of words in Persian electronic texts:

1. In typing Persian texts, typists do not reach a unified way of writing, even by following the grammar of Persian orthography published by the Persian Academy of Language and Literature (PALL).

2. According to the cursive nature of Arabic scripts, two potential forms of writing can be envisaged for a word consisted of at least two morphemes: 'concatenative' where the final letter of a token attaches to the next coming token; and 'non-concatenative' where a blank or Zero-Width-Non-Joiner (ZWNJ) inserts between the tokens. Bear in mind that there exist some letters in Persian which do not join to the next letter such as 'آ – ا' /ā/, 'د' /d/, 'ذ' /z/, 'ر' /r/, 'ز' /z/, 'ژ' /ž/, and 'و' /v/. If the first token ends with any of these letters, then there will be at most two non-concatenative forms. Generally speaking, if a word has n tokens, then the possible number of written forms of that word will be 3^{n-1}. For example, considering the word 'می‌فروختم' /mi + foruxte + ʔam/ 'I have been selling' in which $n = 3$, there will be nine possible forms: 'می^فروخته^ام', 'می فروخته ام', 'میفروختم', *'می^فروختم', 'می فروخته^ام', 'میفروخته^ام', *'می^فروخته ام', 'می^فروخته ام', 'میفروخته ام', and *'می فروختم' (the symbol ^ stands for ZWNJ). The forms with asterisks are basically ill-formed because the final letter of the token 'فروخته' /foruxte/ 'sold' which is Silent Heh 'ه' should not attach to the following token. With this method, the problem of determining the boundaries between Persian words will be reduced to the level of determining the orthographic variations of the morphemes which make up the word. Following Cloeren (1999) if each written token consists of more than one

morpheme, then we will have a multi-unit token (MUT) such as 'بوسه‌ای'/buse + ʔi/ 'a kiss', and 'رفت‌وآمد'/raft + o + ʔāmad/ 'traffic'; and if some tokens, in whole, make up a linguistic unit, we will have a multi-token unit (MTU) such as 'بوسهای'/bushāye/ 'kisses of', and 'رفتوآمد'/rafto ʔāmad/ 'traffic'. As a result, a word in the Persian text can be considered as a MUT or MTU. MTUs are mostly normalized according to the orthographic variations.

Morphophonemic processes in the word formation are usually reflected in orthographic representation, as some MTUs show. For example, the allograph of Heh 'ـهـ' /h/inserts between the two vowels/e/in the word 'بهم'/behem/'to each other', and Alef 'ا' inserts between the vowel graphemes Silent Heh 'ه' and Yeh 'ی' in the word 'بوسه‌ای'.

It is worth mentioning that token concatenation will result in an ill-formed MTU form, if the last allograph of the token to which the next token concatenates happens to be Silent Heh 'ه'/h/; or if one of the concatenated tokens happens to be the conjunctor 'و' /va/ 'and' in structural template of the form. 'XوX' (X denotes a morpheme or allomorph). Therefore, we need a standard to normalize MTU forms.

In recent years, a strategy was prescribed by the PALL for the grammar of the CP orthography which concentrates upon independence of tokens from each other within the MTUs by using ZWNJ between tokens, while the whole MTU is surrounded by a blank to keep its unity in the running text. To generalize MTU evaluation, the MTUs of different types which are inflectional, derivational, and specific compound were obtained from Peykare by using an automatic substring search in input texts. Results showed that contrary to the uniform treatment of the PALL strategy for different types of MTUs, when prefixal and compounding tokens tend to be transparent and separate from the neighboring tokens except for derivational 'به' /be/ (adverbial '-ly'), suffixal and enclitic tokens are prone to concatenate to the neighboring tokens which result in the opacity of the morpheme boundary.

The systematic statistical tendency observed in the MTU forms of Peykare can be explained by the following descriptive generalization to evaluate and normalize MTUs' orthographic variation based on PALL:

"Orthographic words cannot include a blank as a word boundary, and this requirement is enforced by (a) or (b), except when the result is inconsistent with (c), (d), (e) or (f)":

(a) ZWNJ inserts between a prefixal or a compounding token and a following one.

(b) Suffixal or enclitic tokens should concatenate to the neighboring tokens.

(c) No token ending with Silent Heh 'ه' should concatenate to the following one.

(d) The conjunctor 'و' /va/ 'and' should not concatenate to its preceding token ending with a joining letter.

(e) Token concatenation is prohibited when suffix, enclitic, or compounding token begins with Alef 'ا' or Alef with Mad above 'آ'.

(f) Some exceptional derivations and compounds do not obey (a) or (b) for orthographic, aesthetic or any other reason."

One issue close to tokenization in CP text processing is lemmatization. Lemmatizing MUT inflected words to find lexemes or stems of the word formation

is useful in many respects such as morphological analysis, word stemming, and NLP applications like information retrieval. Although a large number of words may occur with a very small frequency as a result of Zipf's law (Manning and Schütze 1999), coverage of a wide domain of registers in a language resource results in the richness of lexicon, thus it causes the reduction of the bias error to some extent. To gain a practical knowledge to deal with the problem of automatic lemmatization of Peykare texts (Mosavi-Miangah 2006), we firstly decided to lemmatize the TC texts in which each word is provided by an EAGLES-based hierarchical tag, as will be described in the next section. To this end, an automatic process of stripping off clitics and affixes from the inflectional MUT words for 2,990 original non-lemmatized text files of the TC resulted in the same number of lemmatized text files. In the second step the TC lexicon was captured by lexical ordering of all lemmatized TC texts.

5 Linguistic annotations

To annotate Peykare with POS tags, we collected a small corpus as a training data set of the TC for automatic POS tagging. This sub-corpus consists of Ettela'at and Hamshahri newspapers from the years 1999–2000, dissertations, books, magazines, blogs, written and spoken texts were collected randomly from 68 different subjects pertaining to different registers to cover varieties of lexical and grammatical structures. The size of the TC has reached 10,612,187 tokens, in which it decreased to 9,781,809 words after tagging some MTUs by means of one specific tag and considering each MTU as a word. This represents an 8% reduction in the corpus size. This reduction means any Persian tokenizer should find a satisfactory algorithm to deal with about 8% of the size of a given text for MTU resolution. This size is computed without taking into account complex predicates such as complex infinitives and verbs. The TC consists of 2,990 text files, each with at least one subject. DBRG at the University of Tehran has provided 2.6 million words of the first version of the TC called the 'Hamshahri Corpus'.

In sum, the TC contains 8,856 subtexts mostly about politics. The chosen subjects are based on the classification made by media. The TC dictionary contains 146,665 non-lemmatized entries in which 27,317 entries are non-linguistic symbols, Arabic and English strings of letters.

5.1 EAGLES-based tagset

The EAGLES guidelines (Leech and Wilson 1999) have been used to mark-up grammar of the texts for the European languages; however, they are also used for tagging the texts of non-European languages as well such as Japanese (Kawata 2001) and Arabic (Khoja et. al. 2001). We have also benefited from these guidelines for Persian because besides being a member of the IE family, its inflectional morphology is rich enough for nouns and verbs at least in comparison with English. As a sample, in Table 3 we have shown three categories of Persian based on the

EAGLES guidelines. Tags are defined on the basis of the major categories (POS) and attributes. It should be added that in our task, contrary to EAGLES which only has 13 major categories including adposition, we have defined two separate categories for preposition and post-position instead of adposition so the major categories added up to 14. The reason for dividing adpositions into the two categories is that the only postposition of CP is 'را' /rā/ (which functions as a definite marker) and it is more active than prepositions in fusing with other major categories. Ultimately, we have 14 tags for the major categories, 52 tags for the recommended attributes, 25 tags for the generic attributes, and 18 tags for the language-specific attributes which add up to 109 tags.

The tags have been given names on a mnemonic framework so that the value of the categories can be easily defined. The structure of the whole given name of a tag is hierarchical, i.e. the major category, the recommended attributes, the generic attributes, and the language specific attributes are represented respectively and they have been separated by commas. The predictable values of some attributes have not been specified. The semantic features, such as the generic attributes of nouns, have been used for distinguishing homophones. For example 'دوشنبه' /došanbe/ could be 'Monday' with the tag name N,PR,SING,DAY (which stands for Noun, PRoper, SINGular, and DAY); or could be 'Dushanbe' with the tag name N,PR,SING,LOC (which stands for Noun, PRoper, SINGular, and LOCation).

The total number of hierarchical tag names of the TC has reached 606 tags. For the main verbs of Persian, a dichotomy of copulative and non-copulative was defined. Mood has been specified merely for the non-copulative verbs; while the copulative verbs are, by default, indicative mood and present tense. Person and number of the subjects of verbs are specified by the numbers 1–6 in which the numbers 1–3 show the first, second, and third person singular, and the numbers 4–6 show the first, second, and third person plural.

The copulative verbs are always accompanied with the indicative mood of the verb /budan/ which is added as an enclitic to a non-verbal element, usually a noun, an adjective, an adverb, a pronoun, or a prepositional phrase. The non-verbal elements of the copulative verbs are tagged with NC,AJC,ADVC,PROC and PC, respectively. The tag SIM has been uniquely used for the copulative simple verb /ʔast/. For example 'خوبم' /xub + am/ 'I am good' is a bi-unit token of which the major category is a verb and its tag will be: V,COP,PRES,AJC,1.

Because of language-specific attributes for Persian, two kinds of morphemes are added to the tagset to distinguish non-lexical homographs from each other and to prepare the necessary information for the process of lemmatization and semi-automatic construction of the treebank. One class of morphemes consists of attaching functional categories such as enclitics to the end of words; and the other one is morphemes or words fusing with the host words and forming a compound tag with at least two tags of the major categories. The most important characteristic of the fused words is that at the morpheme boundary morphophonemic processes are usually involved. For example, 'کو' /ku/ 'that (s)he' appears in literary texts and it is a word which is made from fusing the conjunctor 'که' /ke/ 'that' and the third person singular pronoun 'او' /u/ '(s)he'; so its tag would be: CONJ,PRO,PERS,SING,3.

Table 3 Three EAGLES-based categories for CP

Obligatory attributes	Recommended attributes	Special extensions	
		Generic attributes	Language-specific attributes
1. Noun (N)	*Type*	*Semantic features*	*Enclitics*
	1. COM	1. LOC	1. Pronominals: 1, 2, 3, 4, 5, 6
	2. PR	2. TIME	2. YEH
	Number	3. DAY	3. EZ
	1. SING	4. MON	*Fused with*
	2. PL	5. SEAS	1. POSTP
		6. INFI	2. CONJ
		7. NEG	
		8. SURN	
		9. ACR	
		10. VOC	
2. Verb (V)	*Mood*		*Enclitics*
	1. SUB		1. Pronominals: 1, 2, 3, 4, 5, 6
	2. IMP		*Fused with*
	3. COPR		1. CONJ
	4. PASTP		*Polarity*
	Tense		1. POS
	1. PRES		2. NEG
	2. PA		
	3. FUT		
	4. PERF		
	5. IMPERF		
	6. EIMPERF		
	Copulative		
	1. SIM		
	2. AJC		
	3. ADVC		
	4. NC		
	5. PC		
	6. PNC		
	7. PROC		
	Status		
	1. AUX		
	Person and number: 1, 2, 3, 4, 5, 6		

Table 3 continued

Obligatory attributes	Recommended attributes	Special extensions	
		Generic attributes	Language-specific attributes
3. Adjective (AJ)	*Degree*		*Enclitics*
	1. SIM		1. Pronominals: 1, 2, 3, 4, 5, 6
	2. COMP		2. YEH
	3. SUP		3. EZ
			Fused with
			1. POSTP
			2. CONJ

The enclitics consist of pronominal enclitics, YEH, and Ezafe morphemes. The pronominal enclitics have different syntactic functions, such as subjective, objective, possessive, impersonal, and partative. Each of these enclitics attaches to certain major categories (Megerdoomian 2000). These pronouns are inflected according to person and number. The syntactic functions of such enclitics have not been specified in the full name of a tag. For example, 'خوردمش' /xord + am + aš/ 'I ate it' has been specified by the tag V,PA,SIM,1,3 such that the number '1' is the personal pronoun with subjective function belonging to the recommended attributes, and the number '3' shows the third person singular with objective function belonging to the language-specific attribute.

The morpheme 'ی' /i/ with the tag YEH represents either indefiniteness or relativization of a noun phrase. In either case, it attaches to a noun or the farthest modifier of a noun on which the relativizer conjunctor/ke/will appear. It should be added that the similarity between the pronominal enclitics and the enclitic YEH is that their presence means the end of a syntactic phrase is reached; but, they are in complementary distribution.

Ezafe as another enclitic will be described in Sect. 6. Its difference with the two previous enclitics is that the presence of Ezafe does not determine the end of the syntactic phrase is reached.

The advantage of the EAGLES guidelines in tagging a subcorpus of Peykare can be judged on the basis of the tagset size. The tagset size of a language resource largely depends on the goal of the tagset designer to provide a distinction for all classes of words having a distinct grammatical behavior (Marcus et al. 1993), the inflectionality of the language, and the orthographic representation of words. The POS tagsets developed for English corpora have different sizes according to the different strategies adopted for POS tagging: the Brown corpus with 87 simple tags; the LOB with 135 tags; the UCREL with 165 tags; the LLC with 197 tags; the Penn Treebank with 48 tags; and BNC with 138 tags. Hajič (2000) has shown that the tagset size for highly inflective and agglutinative languages can reach 3,000–5,000 tags; as a result, increasing the degree of inflectionality of a language makes the

tagset size bigger. Since the morphology in Persian is agglutinative and somewhere between highly inflective languages like Arabic and Czech, and less inflective like English, for each tag in the TC there exist information about POS classes and details about inflections, Ezafe, and semantic features to have a feature structure centralized to the POS class. Representing the language inflectionality in orthographic words, in English the possessive construction is represented by a minimal NP consisting of two separate simple orthographic words: such as 'your book'; while in Persian it can be represented by a minimal NP equal to one MUT word like 'کتابت' /ketāb + at / [noun +possessive enclitic] 'your book'. This example shows that the number of hierarchical tags in Persian starting with noun as a POS class must be larger compared to English. Therefore, more and more noun-initial tag names will be added to the Persian tagset if other nominal features like number, indefinite marker, and Ezafe are added to the list of Persian nominal enclitics or suffixes.

Before focusing on the process of tagging in Peykare, it is interesting to determine the advantages of tagging in Peykare compared to PLDB. In Peykare we have used the EAGLES guidelines to standardize tags while tagging in PLDB (Assi and Abdolhosseini 2000) does not follow any special standards. The most distinctive feature in Peykare is that it has used 14 main categories based on EAGLES in which they are enriched linguistically by adding more information to them and there are hierarchical relationships between the tags; while in PLDB only 44 simple tags were used which cannot represent the complexities of Persian.

5.2 Semi-automatic POS tagging

The Editor tool developed for the TC performs two simultaneous operations: segmenting the input raw text into MUTs and MTUs by using a database for free and bound morphemes; and POS tagging semi-automatically. The semi-automatic POS tagging process is as follows:

- Four linguistic graduate students, as annotators, trained to tag words of the very first input raw text files manually by means of the corpus tools.
- Editor tool was programmed to compute the frequency distribution of different POS tags for each word and to update it as the process of tagging continues. As a result, the first version of the tagged text is derived automatically from allocating the most frequent tag to each word (see Voutilainen (1999) for a contextual probabilistic tagging).
- Annotators corrected wrong tags on the Editor tool and proofread the tagged text. The result was called the second version of the POS tagged text.
- UEPRJ software developed for final correction of the tagged text due largely to inter-annotator inconsistencies (Marcus et al. 1993).

UEPRJ is a powerful tool for data search and correction which provides simultaneous access to frequency vocabulary, word tags, and tagged text files via related databases architecture. For example, the word 'آن' /ān/ 'that' has five tags in the TC depending on the linguistic context it appears in. UEPRJ can be used to list

Table 4 Relative frequency of POS tags after lemmatization

POS tag	Tag name	Relative frequency (%)
N	Noun	39.74
P	Preposition	11.25
PUNC	Punctuation	10.27
AJ	Adjective	9.27
V	Verb	8.89
CONJ	Conjunction	8.48
NUM	Number	3.13
PRO	Pronoun	2.58
DET	Determiner	2.50
ADV	Adverb	1.84
POSTP	Postposition	1.47
RES	Residual	0.37
UCL	Classifier	0.21
INT	Interjection	0.01

each tag accompanying the absolute frequency of the word which has occurred. If correction is needed, the annotator can select the target word with one of its tags and review it in the specific sentences of the source tagged texts it occurred in. In the following, the linguistic description of the five tags of the word/ān/with their corresponding absolute frequencies, and sample sentences have come:

Singular demonstrative pronoun (41,265): 'آن را خوردم' /ān rā xordam/ 'I ate it'
Demonstrative determiner (15,345): 'آن کتاب را خواندم' /ān ketāb rā xāndam/ 'I read that book'
Noun, singular common Ezafe (272): 'از آن شماست' /az ān e šomāst/ 'It is yours'
Noun, singular common (25): 'هر آن ممکنه' /har ān momkene/ 'It is possible at any moment'
Noun, singular Proper (10): 'آن ابرامسون' /ān ābrāmson/ 'Ann Abaramson'

The Editor allocates the tag of singular demonstrative pronoun to the word /ān/ automatically, as long as it is the most frequent tag among the five tags. The annotator can change the automatically allocated tag manually only when the mistake is found after proofreading the tagged text. It is important to mention that most of the wrong tags we found in the two last stages are homographs, homophones, proper nouns and Ezafe markers.

After lemmatization of Peykare the number of hierarchical tags sank to 131 tags from 606 tags which is about a 78% reduction as represented in Table 4. In addition, 39% of the word tokens are nouns, which is 2% more than the finding of Hudson (1994) for English corpora as he claimed the generalization that about 37% of word tokens are nouns for any reasonably large body of written English such as LOB and Brown. We believe that the 2% difference for the Persian corpus is due largely to two reasons: highly frequent usage of the Latinized equivalents of English transliterated into Persian as loanwords in scientific texts; and considering complex

verbs made of a non-verbal word (mostly nominal) or phrases and light verbs as separate words and not MTU.

6 Ezafe construction

Ezafe is an enclitic pronounced /e/ to disambiguate the boundary of a syntactic phrase and a linking element to join the head of a phrase to its modifiers, found in the IE languages like Persian and Pashto (Samvelian 2007). This construction has been studied mostly in the framework of Chomsky's GB theory (Ghomeshi 1996); but the scope of Ezafe defined in this paper is equal to or less than the scope defined in theoretical linguistics. For example, the whole phrase of Ezafe construction for the phrase 'کتاب دانش‌آموز زرنگ' $[_{NP}[_{\overline{N}}[_{N}\text{ketāb}]][_{EZ}\text{e}][_{NP}[_{\overline{N}}[_{N}\text{dāneš-āmuz}]][_{EZ}\text{e}][_{AJ}\text{zerang}]]]$ 'the book of the clever student' is considered the same for both theoretical and text processing viewpoints. But, the phrase 'کتاب آن دانش‌آموز زرنگ' $[_{NP}[_{\overline{N}}[_{N}\text{ketāb}]][_{EZ}\text{e}][_{NP}[_{DET}\text{ān}][_{\overline{N}}[_{N}\text{dāneš-āmuz}]][_{EZ}\text{e}][_{AJ}\text{zerang}]]]$ 'the book of that clever student' has two Ezafe constructions in text processing namely [N EZ] and [DET N EZ AJ]; while one Ezafe construction is theoretically embedded within another [N EZ [DET N EZ AJ]].

The Ezafe construction can be demarcated by function words like conjunctors, determiners, postposition 'را' /rā/, some prepositions, verbs and non-verbal elements of a complex verb. A frequency counting of POS categories accepting Ezafe shows a descending order as follows: noun (82%), adjective (10%), preposition (5%), determiner (1%), number (1%), adverb (0.7%), conjunctor (0.03%), pronoun (0.2%), and residual (0.02%). Statistics showed that, regardless of CP varieties, 23% of words have accepted Ezafe in TC. Moreover, 87% of these words with frequency of at least 1,000 items had no overt orthographic symbol for Ezafe which means on average about 20% of words in CP text can include words with Ezafe while no orthographic symbol is used to refer to it. As a result, Ezafe recognition of the words with no overt orthographic symbol is a challenging issue for language engineers working on Persian. Having a moderate error rate of recognizing Ezafe will result in a rather poor intelligibility of Persian speech synthesizers and also a poor performance of syntactic parsing in Persian machine translation which result in increasing the error rate of phrase boundary detection.

To investigate the structure of the Ezafe construction, the following regular expression is applied to TC:

$$^{*}.\text{Ezafe}, ^{*}, \text{NOT}(^{*}.\text{Ezafe})$$

This regular expression is defined to match any POS tag sequence consisting of a tag with Ezafe, followed by any number of tags with Ezafe, ending in a tag without Ezafe. A tag sequence found in this way provides useful information about the length of the Ezafe construction and perhaps semantic constraints among tags within the construction. Table 5 shows the result of pattern matching for the most frequent POS tag sequences of the Ezafe construction. The weighted average of length for such POS tag sequences equated to 2.53 tags. As the length of the Ezafe

Table 5 Frequent POS tag sequences of the Ezafe construction

Tag sequence	Length of tag sequence	Relative frequency (%)
N N	2	33.24
N AJ	2	23.58
N N N	3	8.84
N N AJ	3	6.40
N PRO	2	5.42
AJ N	2	5.38
N NUM	2	4.24
N AJ N	3	4.19
N DET	2	2.67
P N	2	2.42
N N N N	4	2.17
N N N AJ	4	1.44

construction increases, the frequency decreases. Note that almost 44% of the Ezafe constructions are 'hapax legomena' i.e. they occurred once; and 78% of them occurred ten times or less which are largely made up of combinations of a noun with other nouns found in registers like scientific articles, official news, and governmental protocols/documents.

Automatic demarcation of the Ezafe construction is a controversial issue in Persian text processing. As the length of a word sequence increases, recognizing the Ezafe construction correctly will be harder in the absence of any orthographic Ezafe enclitic. We hypothesize that semantic features of the words within a word sequence defined by the EAGLES-based tagset could resolve such a problem, before a syntactic–semantic or discourse methodology is tried.

To judge this proposal, two other regular expressions were defined to search for instantiations of the most frequent Ezafe construction pattern i.e. 'Noun-Ezafe Noun', with this regular expression:

$$\text{noun.}^{*}.\text{Ezafe}(\text{noun},^{*}\ \text{AND NOT}(\text{noun.}^{*}.\text{Ezafe}))$$

and its counterpart without Ezafe, i.e. 'Noun Noun', with this regular expression:

$$(\text{noun.}^{*}\ \text{AND NOT}(\text{noun}.^{*}.\text{Ezafe}))(\text{noun}.^{*}\ \text{AND NOT}(\text{noun.}^{*}.\text{Ezafe}))$$

In Tables 6 and 7 the five most frequent sequences obtained by applying the two regular expressions on the TC are represented. It is found that the semantic features of 93% are two-tag sequences for both kinds. The semantic features LOC (LOCal), TIME, DAY, DIR (DIRection), SURN (SURName), ACR (ACRonym), MON (MONth) and YEH provide appropriate cues to detect the Ezafe construction; as a result, the Ezafe construction detection can be improved at least by using semantic features of the words inserted in the lexicon.

Table 6 Characteristics of frequent two-tag sequences fitting into an Ezafe construction

First tag	Second tag	Typical examples	Relative frequency (%)
N,COM,SING,EZ	N,COM,SING	مورد بررسی، محیط زیست	35.48
N,COM,SING,EZ	N,COM,PL	سازمان ملل، جام آزادگان	8.50
N,COM,SING,EZ	N,PR,SING,LOC	ملت ایران، اتحادیه اروپا	8.26
N,COM,SING,EZ	N,PR,SING	زبان فارسی، شبکه اینترنت	6.13
N,COM,PL,EZ	N,COM,SING	حقوق بشر، نمایندگان مجلس	6.09

Table 7 Characteristics of frequent two-tag sequences not fitting into an Ezafe construction

First tag	Second tag	Typical examples	Relative frequency (%)
N,COM,SING	N,COM,SING	بررسی قرار، استفاده قرار	31.48
N,PR,SING	N,PR,SING	رضا حسینی، بیل کلینتون	10.45
N,PR,SING,LOC	N,COM,SING	روسیه اعلام، تهران اعلام	9.26
N,PR,SING,SURN	N,PR,SING	آیت‌الله خامنه‌ای، امام خمینی	7.07
N,COM,PL	N,COM,SING	بررسی‌ها نشان، تحقیقات نشان	6.60

7 Homograph

Homography, in our terminology, refers to one of two or more words that have the same spelling but differ in meaning and pronunciation, and not necessarily belonging to the same family of languages. The differences in pronunciation can make differences in short vowel and/or stress structure of homographs. Here we are more concerned with the Persian homographs made up from adhesion of suffixes and also enclitics to the stem of at least one homographic word which is one of the most critical issues of the Persian POS tagging. We call such homographs 'non-lexical homographs'. In contrast, 'lexical homographs' are found directly in the lexicon like 'sow' in English with two different meanings and pronunciations of /soʊ/ and /saʊ/. Non-lexical homographs can be classified into different classes in terms of the major morpho-syntatic category each homograph belongs to, such that the members of each class obey an exact orthographic and phonological pattern. In this paper, the sporadic homographs having barren patterns and lexical homographs are excluded from our study.

Because of the productive structure of non-lexical homographs, their analysis in Persian texts is a crucial task while building a Persian resource. Based on the experiment on Peykare, Persian non-lexical homographs can be classified into 13 patterns presented in Table 8 with pattern names, examples, POS patterns, and their frequency distribution both in the TC and Peykare. Pattern names are selected according to the enclitics or suffixes added to one of the homographs.

The homograph richness of Peykare was judged in a process of three steps: firstly, the sets of word tokens with more than one major category, representing 13

Table 8 Statistical results of homograph analysis in the TC and Peykar

Pattern name	Typical examples of homographs	Freq. of homographs in TC	POS pattern (absolute frequency in the TC), an example			Relative freq. in Peykare (%)
Verbal 3rd person	رود	207	V,*,3 (179342) /rav + ad/ '(s)he goes'	N,* (40681) /rud/ 'river'		43.07
Nominal indefinite marker	آسمانی	2117	AJ.* (230933) /āsemāni/ 'celestial'	N.*.YEH (58634) /āsemān + i/ 'a/the sky'	ADV.* (3382) /āsemāni/ 'heavenly'	27.03
Preterite 3rd person	برداشت	79	V,PA,SIM,POS,3 (94409) /bardašt/ 'taking'	N,* (28830) /bardaš + t/ 'harvest/understanding'		12.58
Adjectival indefinite marker	خوبی	890	N,* (47637) /xub + i/ 'goodness'	AJ,SIM,YEH (17574) /xub + i/'a/ the good'		8.46
Copulative adjective 2nd person	خوبی	118	N.* (18792) /xub + i/ 'goodness'	V,COP,*,2 (197) /xub + i/ 'you are good'		3.43
Preterite/perfect 1st person	مردم	43	N,* (16821) /mardom/ 'people'	V,PA/PERF,*,1 (3577) /mord + am/ 'I died'		2.2
Adjectival indefinite marker	دیوانه‌ای	53	AJ,SIM,YEH (1650) /divāne + ʔi/ 'a fool'	V,COP,*,2 (90) /divāne + ʔi/ 'you are fool'		1.44
Possessive 2nd person	درست	6	AJ.* (1324) /dorost/'correct, right'	ADV.* (637) /dorost/ 'correctly'	N,*,2 (16) /dars + at/ 'your lesson'	0.64
Copulative noun 1st person	مردم	40	N,* (2341) /mardom/ 'people'	V,COP,*,1 (125) /mard + am/ 'I am a man'		0.5
Copulative adjective 1st person	شوم	33	AJ,* (338) /šum/ 'inauspicious'	V,COP,*,1 (132) /šav + am/ 'I become'		0.37
Adjectival 3rd person	بارش	72	N,* (581) /bār + aš / 'his/her load'	AJ,SIM,3 (269) /bār + eš/ 'raining'		0.16

Table 8 continued

Pattern name	Typical examples of homographs	Freq. of homographs in TC	POS pattern (absolute frequency in the TC), an example		Relative freq. in Peykare (%)
Subjunctive 1st person	رویم	12	V,SUB,POS,1 (132) /ru + yam/ 'I grow'	N,*,1 (30) /ru + yam/ 'my face'	0.09
Adjectival 2nd person	پاکت	9	N.* (189) /pākat/ 'pocket'	AJ,SIM,2 (15) /pāk + at/ 'your clean'	0.03

types of homographs, were identified by searching the TC. Secondly, the identified words of each type were classified into two or three POS patterns, in which a star is used to show a range of values of attributes defined for major categories (see Table 3). For example, the POS pattern N, *, YEH covers the hierarchical tag names such as N,COM,SING,YEH; N,COM,PL,YEH; or N,PR,PL,YEH among others. Lastly, relative frequency of homographs was computed for each type in Peykare.

The analysis of the results shows that the pronominal clitics and YEH are among the main sources of non-lexical homographs in CP. The six most frequent types of homographs involve the contrast between nouns, verbs, and adjectives. Syntactic and lexical semantic features can be used to resolve noun homographs. Hearst (1991) has checked the contextual surrounding of the target noun to disambiguate English noun homographs using large text corpora. Assuming that a homographic word is ambiguous between a nominal (noun or adjective) and verbal category, a null hypothesis might be that if syntactic context convinces us that the word has to accept Ezafe, then verbal homograph will be rejected since verbs do not accept Ezafe. However, if syntactic context does not provide evidence of Ezafe for the homographic word, then ambiguity will remain unsolved.

Another big challenge that the Persian NLP community should deal with is recognizing noun versus adjective. This is very important for applications like machine translation and TTS. From the Persian TTS point of view, this challenge may be more crucial because poor recognition of the first type of homographs will result in wrong pitch accent patterns of sentences.

8 Summary and future work

In this paper, we explained the major issues in building and evaluating written corpora in contemporary Persian on the basis of findings from two resources: a register-diversified corpus called 'Peykare' and a training corpus annotated by the EAGLES-based POS tagset. After defining five linguistic varieties and 24 different registers based on these linguistic varieties, we collected the texts for Peykare to do linguistic analysis including cross-register differences. In tokenization process of Persian which is challenging for corpus designers, we should deal with multi-token units and multi-unit tokens. To this end, we proposed a descriptive generalization to normalize orthographic variations existing in texts. To annotate Peykare, we benefited from the EAGLES guidelines to have tag hierarchies as a result. For the methodology used in the annotation of Peykare, we have used a semi-automatic approach. The Ezafe construction and homographs, which are problem makers in text processing, were discussed.

As for the future work, we will use the tags for automatic treebanking of the TC as the training data for treebanking of Peykare.

Acknowledgments This project was funded by the Higher Council for Informatics of Iran and the University of Tehran under the contract number 190/3554. Masood Ghayoomi was funded by the German research council DFG under the contract number MU 2822/3-1. Our special gratitude also goes to Dr. Ali Darzi at the University of Tehran who cooperated with us in the project and the anonymous reviewers for their helpful comments. However, the responsibility for the content of this study lies with the authors alone.

References

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics, 11*(2), 135–171.

Assi, M., & Abdolhosseini, M. H. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics, 5*(1), 69–81.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1–16.

Biber, D. (1992). Representativeness in corpus design. In G. Sampson & D. McCarthy (Eds.), *Corpus linguistics: Readings in a widening discipline* (pp. 174–197). New York, USA: Continuum.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19*(2), 221–241.

Bijankhan, M. et al. (1994). Farsi spoken language database: FARSDAT. In *Proceedings of the 5th international conference on speech sciences and technology (ICSST), Perth* (Vol. 2, pp. 826–829).

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

Bijankhan, M. et al. (2004). The large Persian speech database. In *Proceedings of the 1st workshop on Persian language and computer, the University of Tehran, Tehran, Iran* (pp. 149–150).

Buckwalter, T. (2005). Issues in Arabic orthography and morphology analysis. In *Proceedings of the workshop on computational approaches to arabic script-based languages in conjunction with COLING 2004*, *Switzerland*.

Cloeren, J. (1999). Tagsets. In H. V. Halteren (Ed.), *Syntactic wordclass tagging*. Dordrecht, The Netherlands: Kluwer.

Douglas, F. M. (2003). The Scottish corpus of texts and speech: Problems of corpus design. *Literary and Linguistic Computing, 18*(1), 23–37.

Ghayoomi, M., & Momtazi, S. (2009). Challenges in developing Persian corpora from online resources. In *Proceedingss of IEEE international conference on Asian language processing, Singapore*.

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A study of corpus development for Persian. *International Journal on Asian Language Processing, 20*(1), 17–33.

Ghomeshi, J. (1996). Projection and inflection: A study of persian phrase structure. Ph.D. thesis, University of Toronto, Toronto, ON.

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the 6th applied natural language processing conference, Washington* (pp. 94–101).

Hearst, M. A. (1991). Noun homograph disambiguation using local context in large text corpora. In *Proceedings of the 7th annual conference of the University of Waterloo, Center for the new OED and text research, Oxford*.

Hodge, C. T. (1957). Some aspects of Persian style. *Language*, *33*(3) Part 1, 355–369.

Hudson, R. (1994). About 37% word-tokens are nouns. *Language, 70*(2), 331–339.

Hussain, S., & Gul, S. (2005). *Road map for localization*. Lahore, Pakistan: Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences.

Kawata, Y. (2001). Towards a reference tagset for Japanese. In *Proceedings of the 6th natural language processing Pacific rim symposium post-conference workshop, Tokyo* (pp. 55–62).

Khoja, S., Garside, R., & Knowles, G. (2001). *A tagset for the morpho-syntactic tagging of Arabic*. Lancaster University, Computing Department. http://archimedes.fas.harvard.edu/mdh/arabic/CL2001.pdf.

Kralik, J., & Šulc, M. (2005). The representativeness of Czeck corpora. *International Journal of Corpus Linguistics, 10*(3), 357–366.

Kučera, K. (2002). The Czech national corpus: Principles, design, and results. *Literary and Linguistic Computing, 17*(2), 245–247.

Leech, G. (2002). The importance of reference corpora. *Donostia*, 2002-10-24/25. www.corpus4u.org/upload/forum/2005060301260076.pdf.

Leech, G., & Wilson, A. (1999). Standards for tagsets. In H. V. Halteren (Ed.), *Syntactic wordclass tagging* (pp. 55–81). Dordrecht, The Netherlands: Kluwer.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: The MIT press.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). *Building a large annotated corpus of english: The penn treebank*. http://citeseer.comp.nus.edu.sg/587575.html.

Megerdoomian, K. (2000). *Persian computational morphology: A unification-based approach*. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-320).

Mosavi-Miangah, T. (2006). Automatic lemmatization of Persian words: Project report. *Journal of Quantitative Linguistics, 13*(1), 1–15.

Muthusamy, Y. K., Cole, R. A., & Oshika, B. T. (1992). The OGI multi-language telephone Speech Corpus. In *Proceedings of the 2nd international conference on spoken language processing (ICSLP)*, *Banff* (pp. 895–898).

Samvelian, P. (2007). A (phrasal) affix analysis of the Persian Ezafe. *Journal of Linguistics, 43*, 605–645.

Sheykhzadegan, J., & Bijankhan, M. (2006). The speech databases of Persian language. In *Proceedings of the 2nd workshop on Persian language and computing, the University of Tehran, Tehran, Iran* (pp. 247–261).

Sinclair, J. (1987). *Corpus creation*. In G. Sampson and D. McCarthy (Eds.), *Corpus linguistics: Readings in a widening discipline*, 2004 (pp. 78–84). New York: Continuum.

Voutilainen, A. (1999). A short history of tagging. In H. V. Halteren (Ed.), *Syntactic wordclass tagging* (pp. 9–19). Dordrecht, The Netherlands: Kluwer.