International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

An Analysis of a Persian Corpus

Ali Arabmofrad
Department of English Language and Literature, Golestan University, Iran
a.arabmofrad@gu.ac.ir

ABSTRACT

In the arena of corpus linguistics, Persian is among the languages that lack a thorough and comprehensive corpus of its contemporary use. Therefore, the present study is a preliminary attempt to construct a corpus for Persian language which is equivalent to that of other languages; for example, BNC which has more than 100 million words (spoken or written). To begin, a corpus of five million words of different contemporary newspaper articles that are considered to be, to a great extent, representative of Persian Standard formal use, were gathered. The corpus was analyzed and an overview of the most frequent single words and multi-word strings was found. Furthermore, it was found that the frequency of some multi word strings are more than that of single words that may be considered as a core vocabulary in Persian signaling the fact that such strings and expressions can play much more important role than conceived before.

KEY WORDS: Corpus; Corpus Analysis; Corpus Application; Persian Corpus

INTRODUCTION

CORPUS

"A corpus is a collection of texts, written or spoken, usually stored in a computer database" (McCarthy, 2004, p.1). This collection of words can be used in linguistic description or language hypothesis verification. The sources for written corpora are usually books, newspapers and magazines and those of spoken corpora are everyday (casual) conversations, phone calls, radio and TV programs , lectures, interviews and so on that are transcribed. A corpus may contain millions of words or it may be very small depending upon the purpose and aspect under investigation.

CORPUS LINGUISTICS

Corpus linguistics explores the actual patterns of language use, e.g. analyzing natural language or seeking how language use varies in different situations (i.e., spoken vs. written, formal vs. informal) by using large collections of both spoken and written natural texts (corpora) stored in a computer (Schmitt, 2002). Furthermore, It should not be confused with computational linguistics which a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty (Uszkoreit, 2010).

The major difference between corpus linguistics and traditional linguistics is that corpus linguistics systematically studies authentic examples of naturally occurring language (Wengao, Huaqing, Hong, 2005) and because of the size of these collections, the use of computers for analysis is imperative. Computers store and analyze millions of words according to word tags and also by the use of concordancing programs (McArthur & McArthur, 1992). "A concordancer usually consists of two programs: A word frequency program and a concordance program to find all instances of a given word in a corpus" (Flowerdew, 2001, p.71).

While Widdowson (2000b) states that corpus linguistics (as well as critical discourse analysis) cannot be regarded as an independent discipline but it's just linguistics applied, Renouf (2005) is more inclined to view corpus linguistics as a discipline because it has become self-reflective and self-critical and has reached the state of maturity of a discipline.

TYPES OF CORPORA AND EXAMPLES

Generally, corpora can be divided into general and specialized ones. While general corpora (like BNC and the Brown Corpus) include language samples from a wide range of registers or genres, and contain more than 1,000,000 words, specialized corpora (e.g., the Michigan Corpus of Academic Spoken English, MICASE) contain more register specific collections (written and/or spoken) and are designed with more specific research goals (Schmitt, 2002). More specifically, one can divide the corpora into native speaker vs. learner, monolingual vs. multilingual, original vs. translation, synchronic vs. diachronic, and plain vs. annotated.

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

According to Nation (2001), corpus research has three essential requirements: an idea or a question to be investigated by means of corpus, the source of data; i.e., a corpus, and the computer programs, e.g., VocabProfile and RANGE (Nation, 2001), the Word Smith Tools and Word List (Scott, 2001).

The creation of the Brown Corpus as the first computer corpus goes back to the early 1960s (Meyer, 2002). Some of the more well-known available corpora include the Cambridge International Corpus, the British National Corpus (BNC), the Brown Corpus, the Lancaster/Oslo-Bergen (LOB) Corpus and the Helsinki Corpus of English Texts. From such corpora one can obtain information about the frequency of occurrence, lexical co-occurrence patterns, and (unnoticed) patterns of language use (Schmitt, 2002).

Some of the main English corpora developed so far are geographical varieties (e.g., the Brown corpus; the Lancaster-Oslo-Bergen corpus or LOB) which were collections of written data, spoken language corpora (e.g., the London-Lund Corpus, the Corpus of Spoken American English), mixed corpora (e.g., the British National Corpus which is a collection of 100 million words, 90% of which are from written sources and the rest from spoken ones), historical varieties (e.g., the Helsinki corpus of English texts which has samples of different dialects of three historical periods (Old, Middle, and Early Modern English), child and learner varieties (e.g., CHILDS), genre- and topic-specific corpora (e.g., HCRC map task corpus which consists of 128 transcribed performances of map-reading tasks), multilingual corpora (e.g., the European Corpus Initiative (ECI) that has produced a multilingual corpus of over 98 million words, covering most of the major European languages, as well as Turkish, Japanese, Russian, Chinese, Malay and so on).

ISSUES IN CORPUS DESIGN AND ANALYSIS

The first featureof a corpus design is the size of the corpus which depends upon the purpose for which the corpus has been collected. For example, to provide information about the frequency of words (especially for knowing the less frequent lexical items) the corpus should contain millions of words while for different grammatical constructions the size can be smaller "since there are far fewer different grammatical constructions than lexical items, and therefore they tend to recur much more frequently in comparison" (Schmitt, 2002, p.96).

Another feature in designing a corpus is that it should be as representative as possible of the type of the language included in it. This representativeness may be accomplished by including "different registers (fiction, non-fiction, casual conversation, service encounters, broadcast speech), discourse modes (monologic, dialogic, multi-party interactive) and topics (national versus local news, arts versus sciences, etc.), and even demographics of the speakers or writes (nationality, gender, age, education level, social class, native language/dialect)" (Schmitt, 2002, p.96).

Biber. Conrad and Reppen (1998, p. 4) identify four characteristics of a corpus based analysis of language:

. It is empirical, analyzing the actual patterns of use in natural texts.
. It utilizes a large and principled collection of natural texts, known as 'corpus', as the basis of analysis.
. It makes extensive use of computers for analysis, using automatic and interactive techniques.
. It depends on both qualitative and quantitative analytical techniques

The analysis of the corpora may be done both qualitatively and quantitatively. Quantitative analysis, accomplished by computers, refers to macro-level characteristics and involves statistical techniques and mechanical tasks. Qualitative analysis, on the other hand, deals with micro level characteristics done by humans who decides upon the type of information to search, how to extract the information and how to interpret them (Schmitt, 2002).

CORPUS APPLICATION

At the first glance, the main uses of corpus linguistics may be seen as the frequency of words of a language and the number of words needed to teach, the differences that exist in speaking and writing, the contexts of use and collocation, grammatical patterns, etc.

Meyer (2002) enumerates some of the applications of corpus-based research in linguistics: Studying a particular grammatical structure regarding its form, usage, etc in detail (Grammatical studies of specific linguistic constructions) and the usage of many different grammatical structures (Reference grammars), teaching vocabulary (see McCarten, 2007), developing dictionaries and improving word information contained in them (lexicography), creating corpora regarding language variation balanced considering age, gender, social class and regional dialects (language variation), studying earlier dialects of a particular language (historical linguistics), Natural language processing, first- and second-language acquisition, e.g., CHILDES (Child Language Data Exchange) and in language pedagogy (e.g., “data-driven learning” (Johns 1994 and Hadley 1997) advocates the students' studying corpora to learn about English and in contrastive analysis and translation theory.

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

PARALLEL CORPORA

Although, the first collected corpora were largely monolingual, the new advancement in the area of computer technology and the enormous amount of information on the Internet has made it easy to develop multilingual or parallel corpora. Such corpora can consist of single or several different translations of the original text. From parallel corpora one can obtain information about translation equivalents, collocational and phraseological units, machine translation, multilingual information retrieval (Scannell, 2003) and vocabulary building (Chujo, Utiyama ,Nishigaki, 2004). The English–Norwegian Parallel Corpus is an example of the earlier parallel corpora (to know more about different projects in parallel corpora, see Scannell, 2003).

THIS STUDY

In recent years, the corpora of many languages of the world have been developed , but there appears to be no such corpus in Persian which is widely spoken in Iran, Afghanistan, Tajikistan, Uzbekistan and to some extent in Iraq, Bahrain, and Oman. In Afghanistan Persian is known as Dari or Dari-Persian, while in Tajikistan it's known as Tajiki and the official language of Iran is sometimes called Farsi in English and other languages.

There are many questions, yet unexplored, that can be answered by corpus studies in Persian language. At the moment we do not know the most frequent words, the most frequent multi word strings, the most frequent phrasal verbs in Persian, the tenses people use most frequently, prepositions following particular verbs, collocation patterns, words used in more formal situations, and in more informal ones, the use of idiomatic expressions, and the number of words a learner must know in order to participate in everyday conversation in Persian.

Therefore, the present study is a preliminary attempt to create and develop Persian National Corpus (PNC) that is going to be comparable to the BNC. Because the area of sample selection is too broad and time consuming to deal with within a single study, we began the compilation of the Persian corpus from written sources collected from newspaper articles.

DATA

A five-million- word corpus was created by extracting the texts of newspaper articles. The data were all collected from the electronic version of these newspapers. In order to ensure the representativeness of all the newspapers published in Iran a total of 9 Newspapers were selected out of 50 identified national newspapers. Furthermore, as shown in see table 1, to ensure the representativeness of subject area, different sections of the mentioned newspapers were enumerated and the following list was selected: science and technology, politics in Iran and in the world, society, economic, sports, art and cinema.

Table 1
The Proportion of the Main Categories

Category Name in Persian	Category Name	Proportion
اقتصادی	Economics	21.01%
اجتماعی	Social	9.55%
هنری-فرهنگی- سینمایی	Culture-Art-Cinema	7.62%
علمی	Science	3.61%
ورزشی	Sports	25.4%
سیاسیو خارجیداخلی	Politics (Domestic & International)	32.78%

Each category consists of different subcategories. For instance, economics category consists of 17 subcategories some of which are banks and insurance, business, bourse, car and computer market. In order to minimize the effect of some special events on the data, they were collected from different points in time.

SOFTWARE

A text-analyzer software similar to Wordsmith Tools for English was developed to retrieve the strings of characters and spaces (words) and to give a count for their occurrence in the data. We set the number of words for recurrent strings at 20. In order to have more speed in obtaining data from software, the single words or multiword strings whose number of occurrence were less than 30 were deliberately not considered (i.e., the cut-off point for occurrence was 30). As stated by McCarthy (2006), the important and predictable point about such strings is that they may lack syntactic or semantic integrity. This is because of computer's lack of sensitivity towards meaningful

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

and meaningless strings. The general procedure to analyze the five-million word corpus collected from different newspaper texts at different points in time was to generate lists of single-, two-, three-, four-, and five- word sequences. The lists are rank-ordered according to their frequency of occurrence. As the number of the words of clusters increases, the fall-off between clusters becomes more and more sharper. Figure 1 shows comparatively the distribution of clusters in excess of 20 occurrences.

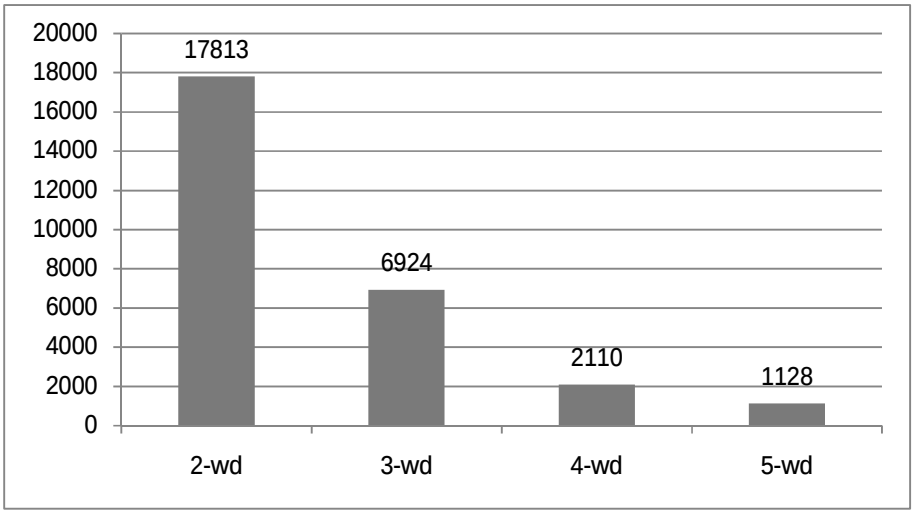

Figure 1. Distribution of clusters in excess of 20 occurrences

RESULTS

Table 2 to 6 show the most frequent items in single words and 2-5 clusters

Table 2
Top 20 Single Words

	Word	Gloss	Meaning	Grammatical category	Freq.
1	و	/væ/	and	conjunction	165613
2	در	/dær/	in	preposition	146983
3	به	/be/	to	preposition	120784
4	از	/æz/	of/from	preposition	91505
5	این	/i:n/	this	determiner	75851
6	را	/ra/	object indicator	OM	61968
7	با	/ba/	by/with	preposition	60441
8	که	/ke/	that	conjunction	57321
9	است	/æst/	is	linking verb	52677
10	براي	/bæraje/	for	preposition	18146
11	آن	/an/	it/that	pronoun/determiner	17816
12	خود	/xod/	self	anaphor	17713
13	کرد	/kærd/	did	auxiliary verb	17375
14	تیم	/tim/	team	noun	16696
15	که	/ke/	that	conjunction	16660
16	شده	/shode/	has become	auxiliary verb	16308
17	شد	/shod/	became	auxiliary verb	16229
18	گفت	/goft/	said	verb	15062
19	ایران	/Iran/	Iran	noun (proper name)	14348
20	بر	/bær/	on	preposition	14267

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

Table 3
Top 20 Two-Word Clusters

	Word	Gloss	Meaning	Freq.
1	دراین	/dær in/	in this	12296
2	رابه	/ra be/	OM to	8698
3	استکه	/æstke/	is that	8011
4	کهدر	/kedær/	that in	6728
5	را در	/radar/	OM in	6205
6	خود را	/kodra/	Self OM	5848
7	شده است	/shodeæst/	has been	5731
8	و در	/vædær/	and in	5716
9	است و	/æstvæ/	is and	5662
10	پس از	/pæsaz/	after	5401
11	از این	/æz in/	from/of this	5388
12	به این	/be in/	to this	4421
13	و به	/væ be/	and to/for	4262
14	به گزارش	/be gozareshe/	according to	4058
15	به عنوان	/be onvane/	as	3966
16	در سال	/dærsal/	each/in the year	3364
17	را از	/raæz/	OM of	3191
18	و با	/væba/	and with	3150
19	که به	/ke be/	conjunction to	3056
20	اشاره به	/eshare be/	referring to	2893

Table 4
Top 20 Three-Word Clusters

	Word	Gloss	Meaning	Freq.
1	با اشاره به	/baeshare be/	referring to	2828
2	با توجه به	/be tavæjjoh be/	considering	2083
3	در حال حاضر	/dær hale hazer/	at present	1452
4	در پاسخ به	/dærpasokh be/	in answer to	1000
5	با بیان این	/babayan in/	stating	984
6	این است که	/in æstke/	this is that	981
7	به گزارش ایسنا	/be gozareshisna/	according to ISNA	958
8	در این زمینه	/ dær in zæmine/	in the filed of	901
9	خود را به	/khodra be/	self OM to	874
10	است که در	/æstkedær/	is that in	845
11	خود را در	/khodradær/	self OM in	814
12	آموزش و پرورش	/amuzeshvæpazhuhesh/	education and training	704
13	جمهوري اسلامي ايران	/jomhuriyeeslamiyeiran/	Islamic Republic of Iran	653
14	با بیان اینکه	/babayaneinke/	stating that	637
15	مجلس شوراي اسلامي	/majleseshorayeeslami/	Islamic Parliament	626
16	کرد و گفت	/kærdvægoft/	did and said	623
17	به نقل از	/benæghlæz/	quoting	608
18	در حالي که	/dærhalike/	while	602
19	پیش از این	/pishæz in/	before	579
20	در این باره	/dær in bare/	in this regard	575

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

Table 5
Top 20 Four-Word Clusters

	Word	Gloss	Meaning	Freq.
1	در گفت وگو با	/ dærgoftoguba/	interviewing with	777
2	اين در حالي است	/in dærhalistke/	while	622
3	وي با اشاره به	/vey baeshare be/	he referring to	547
4	در حالي است که	/dærhaliæstke/	while	524
5	با بيان اين که	/babæyaneinke/	stating/saying	479
6				
7	به نظر مي رسد			
8	با بيان اين مطلب	/babæyane in mætlab /	stating this point	363
9	در پاسخ به اين	/dærpasokh be in/	in reply to	326
10	و با توجه به	/væbatavæjjoh be/	and considering	319
11	با اشاره به اين	/baeshare be in/	referring to	242
12	با اشاره به اينکه	/ baeshare be inke/	referring to	233
13	درگفتگو با ايسنا	/ dærgoftogubaisna/	in interview with ISNA	215
14	خبر داد و گفت	/khæbær dad vægoft/	announced and said	211
15	با اعلام اين خبر	/baelame in khæbær/	announcing the news	199
16	مجمع تشخيص مصلحت نظام	/majmaetæshkhisemæslehætenezam/	Expediency Council	186
17	را به خود اختصاص	/ra be khodekhtesas/	OM assigned to him/her/itself	174
18	اشاره کرد و گفت	/esharekærdvægoft/	referred and said	166
19	و به همين دليل	/væ be hæmindælil/	and for this reason	160
20	با توجه به اين	/batævajjoh be in/	considering this	157
21	در نظر گرفته شده	/dærnæzærgerefteshode/	is considered	146

Table 6
Top 20 Five-Word Clusters

	Word	Gloss	Meaning	Freq.
1	اين در حالي است که	/in dærhaliæstke/	while	646
2	با اشاره به اين که	/baeshare be in ke/	with reference to this that	166
3	امنيت ملي و سياست خارجي	/æmniyatemelli va siasætekhareji/	national security and foreign policy	134
4	در ادامه با اشاره به	/dæredame be eshare be/	furthermore referring	130
5	جامعه مدرسين حوزه علميه قم	/jameeyemodæresinhozeyeelmiyeyeghom/		124
6	ملي و سياست خارجي مجلس	/mellivæsissatekharejiyemajles/	national and parliament's foreign policy	116
7	در پاسخ به اين سؤال	/dærpasokh be in soal/	in answering to	109 102
8	سازمان بورس و اوراق بهادار	/sazmane burs væoraghebæhadar/	stock exchange	103
9	با بيان اين مطلب افزود	/babæyane in mætlæbæfzud/	stated this point and added	100

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

10	با تأکید بر این که	/batækidbærinke/	emphasizing	94
11	پاسخ به این سؤال که	/pasokh be in soalke/	answer to this question	86
12	وی با بیان این که	/vey babæyaneinke/	he stating that	83
13	کمیسیون امنیت ملي و سیاست	/komisiyuneæmniyatemelivæsiasæte/	national security committee and …policy	83
14	با اعلام این خبر گفت	/baelame in khæbargoft/	announcing the news said	80
15	با توجه به این که	/batævajoh be inke/	considering	75
16	به گزارش پایگاه اطلاع رسانی	/begozareshepaygaheettelaræsani/	according to information station	68
17	شرکت مهندسي و توسعه گاز	/sherkætemohændesivætoseeyegaz/	gas development co.	68
18	انتخابات دهمین دوره ریاست جمهوري	/entekhabatedæhomindoreyeriyasætjomhuri/	the tenth presidential election	67
19	غیر از این است که	/gheiræz in æstke/	apart from this	66
20	عضو شوراي مرکزي سازمان مجاهدین	/ozveshorayemærkæzi-ye sazmanemojahedin/	member of the central council of mojahedin organization	65

GENERAL ANALYSIS

Table 2 shows the most frequent single words in the five-million written corpus. As seen, while the first 20 words account almost for 37% of the words, the first five words proportion is 21% and the first word va (and) accounts for 5.81% of the whole corpus. Of twenty most frequent words, 16 are function words and just four are content words. Among these words, 6 of which are prepositions (16.4 % of the corpus), 5 are verbs (4.39 % of the corpus) and just two of them, i.e., Iran and Team are nouns (1.12 % of the corpus).

CONTENT WORDS

The most common verbs are æst (is) which functions both as the main verb and auxilliary, shod (past tense of the verb shodæn which means become), shode (pastparticiple of the verbshodan) and goft (said). The reason for the high frequency of the first four verbs (totally 3.60) is that in Persian they may function not only as the main verb but also as an auxiliary to make compound verbs. In addition they are used to make different tenses (present, past, present perfect, past perfect and future). In Table 7, there are some instances of the verb goftand kærd.

Table 7

Instances of applications of the verb goft and kærd

	Word	Gloss	Meaning	Tense
گفت	گفت	goft	said	past
	می شود گفت	mishævædgoft	it can be said	present
	خواهد گفت	khahædgoft	will say	future
کرد	آغازکرد	aghazkærd	began	past
	بیان میکند	bayanmikonæd	states	present
	تمام خواهدکرد	tamamkhahædkærd	will end/finish	future
	خاموش کرده است	khamushkærdeæst	turned off	present perfect
	نگاه کرده بود	negahkærde bud	looked	Past perfect
	می گرید	migeryæd	cried	present continuous
	خنده کرد	khændekærd	laughed	past

International Journal of Modern Language Teaching and Learning
Available online at www.ijmltl.com. Vol. 2, Issue 1, 2017, pp.32- 40
ISSN: 2367-9328

Contrary to prepositions, verbs and nouns, there is no trace of any adjective or adverb in the first 100 frequent words which indicated the less application of these two categories.

CLUSTERS

As stated by McCarthy (2006), the important and predictable point about such strings is that they may lack syntactic or semantic integrity. This is because of computer's lack of sensitivity in distinguishing meaningful and meaningless strings. The most frequent cluster in this corpus is dær in (in this). Compared to frequencies of single words, there are just 32 single words that occur more frequently than this two-word cluster. Surprisingly, its occurrence (12296) is more than the words such as hæm (also), amma (but), keshvær (country), ma (we), bæraye (for), daræd (has), vey (he) and bayæd (should/must). Figures 2 and 3 compare some other two- three- four- and five-word string frequencies with those of some single words.

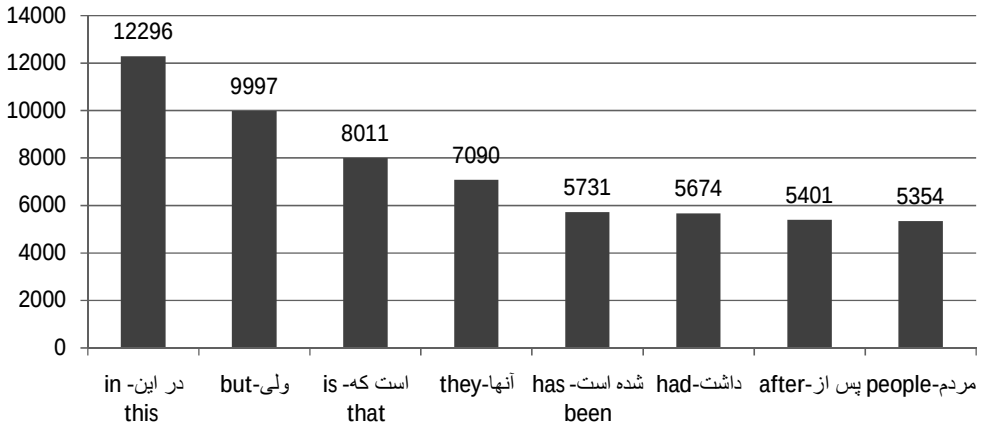

Figure 2. Two-word clusters and single words

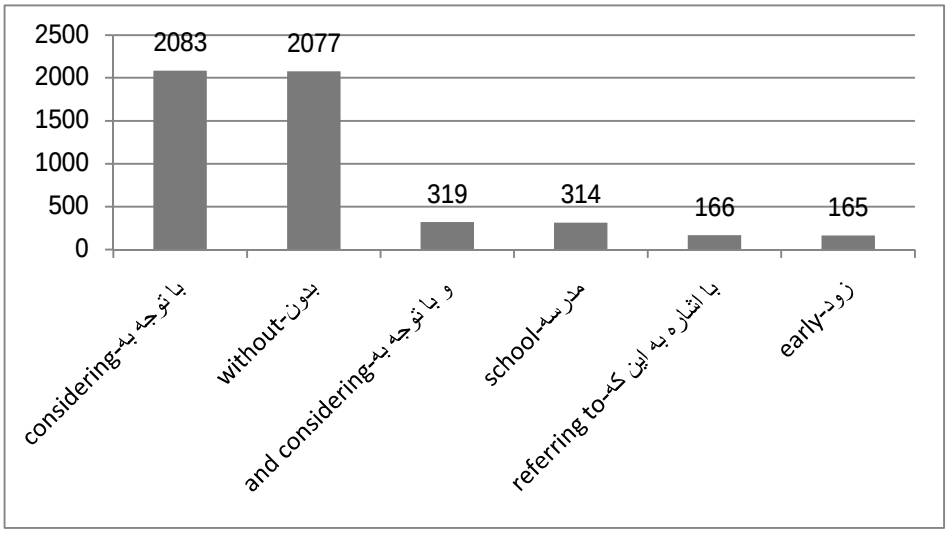

Figure 3. Three-, four- and five-word clusters and single words

The implication of such graphs in the area of language teaching especially in vocabulary teaching is that "the word lists which focus only on single words risk losing sight of the fact that many high frequency clusters are more frequent and central to communication than even very frequent words" (McCarthy, 2006, p.17).

DISCUSSION

Corpus linguistics enables us to uncover the most trivial and hidden words and phrases that are used to lessen the burden of processing in written and spoken mode and it is difficult to introspect on what one says in the absence of corpus evidence (McCarthy, 2006). The findings of such corpuses can be used not only in materials development in teaching Persian but also in developing Persian dictionaries as well. The same study can be conducted to know about the frequency of the words of Persian texts in different fields of study, for example, the most frequent words in engineering texts, and the findings can be used as a guideline to develop special bilingual dictionaries from Persian into other languages.

Conflict of Interest
The authors declare no conflict of interest.

REFERENCES

Biber, D., Conrad, S., &Rappen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Chujo, K., Utiyama, M., Nishigaki, C. (2005). A Japanese-English Parallel Corpus and CALL: A Powerful Tool for Vocabulary Learning. Proceedings of FLEAT5（The 5th Foreign Language Education and Teaching), pp.16-19.

Flowerdew, J. (2001). Concordancing as a tool in course design.In H. Ghadessy, A. Henry & R. L. Roseberry, Small Corpus Studies and ELT.Theory and practice (pp. 71-92). Philadelphia: John Benjamins.

Hadley, Gregory (1997) Sensing the Winds of Change: An Introduction to Data-riven Learning.Retrieved on February 12, 2010, from http://web.bham.ac.uk/johnstf/winds.htm.

Johns, T.F. (1994). 'From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. inOdlin, T. (Ed.) Perspectives on Pedagogical Grammar. (Pp. 293-313). New York: Cambridge University Press.

McArthur & McArthur.(Eds.). (1992). The Oxford Companion to the English Language. Oxford: Oxford University Press.

McCarten (2007).TeachingVocabulary. Cambridge: Cambridge University Press.

McCarthy, (2004). Touchstone: From Corpus to Course Book. Cambridge: Cambridge University Press.

McCarthy, M. (2006).Explorations in corpus linguistics. Cambridge: Cambridge University Press.

Meyer, C. F. (2002). English Corpus Linguistics: An introduction. Cambridge: Cambridge University Press.

Nation, P. (2001). Using small corpora to investigate learner needs: Two vocabulary research tools. In In H. Ghadessy, A. Henry & R. L. Roseberry ,Small Corpus Studies and ELT. Theory and practice (pp. 31-45). Philadelphia: John Benjamins.

Renouf, A. (2005). Corpus Linguistics: past and present, in Wei Naixing, Wenzhong, Li, Pu Jianzhong (eds.), Corpora in Use: In honour of Professor Yang Huizhong.

Scannell, K. P. (2003). Automatic thesaurus generation for minority languages: an Irish example. In Actes de la 10e conf´erence TALN `a Batz-sur-Mer, volume 2, pages 203–212.

Scott, M. (2001).Comparing corpora and identifying key words, collocations, and frequency distributions through WordSmith Tools suite of computer programs.In H. Ghadessy, A. Henry & R. L. Roseberry, Small Corpus Studies and ELT.Theory and practice (pp. 47-67). Philadelphia: John Benjamins.

Schmitt, N. (2002). An introduction to applied linguistics. London: Arnold.

Uszkoreit, H. (2010). What is computational linguistics? Retrieved January 10, 2010, from http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

Wengao, G., Huaqing, H., Kim Hong, K. (2005). Incorporating corpus linguistics into content teaching: the feasibility of using small corpus in Singapore primary maths teaching. Proceedings of the Redesigning Pedagogy: Research, Policy, Practice Conference, Singapore, May-June, 2005.

Widdowson, H. G. (2000b). Onthelimitations of linguistics applied. Applied Linguistics 21(1), 3-25.