Farsi lexical analysis and stop word list

M.R. Davarpanah
Faculty of Education and Psychology, Ferdowsi University of Mashhad, Mashhad, Iran
M. Sanji
Imam Reza University, Mashhad, Iran, and
M. Aramideh
Mashhad Education Organization, Mashhad, Iran

Received 30 March 2009
Revised 24 May 2009
Accepted 11 June 2009

Abstract

Purpose – The purpose of this article is to present an aggregated methodology for construction of the stop word list in Farsi language and generate a generic Farsi stop word list.

Design/methodology/approach – The stop word list is extracted based on: syntactic classes, domain dependent, corpus statistic and expert judgments. Some of the main challenges that arise in the Farsi automatic text processing are outlined as well.

Findings – Results from the techniques are aggregated and a general Farsi stop word list containing 927 words is generated.

Practical implications – The created stop word list can affect the efficiency and effectiveness of retrieval and indexing process in Farsi information retrieval system, moreover, it can play an important role during Farsi text segmentation.

Originality/value – Our stop word extraction algorithm is a promising technique; it could be applied into other languages that they have ambiguities in automatic text segmentation.

Keywords Languages, Information retrieval

Paper type Research paper

Introduction

In information retrieval (IR), a document is traditionally indexed and searched by words. By definition, stop words are very common words that appear in the text frequently but do not carry significant information in terms of information retrieval (Zou *et al.*, 2006). They can affect the retrieval effectiveness because of a very high frequency and tend to diminish the impact of frequency differences among less common words. They can also affect efficiency, resulting in a large amount of unproductive processing (Korfhage, 1997). The removal of stop words can increase the efficiency of the indexing process as stop words can represent 30 to 50 per cent of the tokens in a large text collaboration (Schauble, 1997). Therefore, identifying a list of stop words in order to eliminate them from text processing during indexing as well as before querying is essential to an information retrieval system.

Lots of stop words lists have been developed for English language in the past, which are usually based on frequently statistics of a large corpus (Van Rijisbergen, 1975; Francis and Kucera, 1982; Fox, 1990). Farsi like any other natural language

Library Hi Tech
Vol. 27 No. 3, 2009
pp. 435-449

0737-8831
DOI 10.1108/07378830910988559

contains many stop words that can represent a considerable percentage of the tokens in a text. However, no general and commonly accepted stop word list has been constructed for Farsi language. Research and experimentation in natural language processing (NLP) in Farsi language is fairly new. There are a few studies on Farsi stop words in the literature. Taghva *et al.* (2003a, b) in Nevada University produced a stop list that embodied variant forms of only 12 verbs (saying that each verb had almost as many as 100 variations given its infinitive, imperative and past tens forms). The list they produced also included non-verbal stop words which mounted to 155 words. It is an introductory part of a research on Farsi searching and display technologies and created list is relatively short and incomplete as mentioned by the authors. Falahati Qadimi Fumani and Ramachandra (2008) examined 30 chemistry scientific articles to find out that all punctuation marks, numbers, and English letter combinations can be included in the stop list or not. The study revealed that the omission of all punctuation marks, numbers, etc. will have a negative effect on recall. Since punctuation marks, particularly "dot" and "hyphen" appear in the structure of content-bearing elements and even appear abundantly in titles and abstracts of scientific articles. With the fast growth of online Farsi documents and the rapid increase of Farsi web sites, constructing a general Farsi stop word list with an applicable generating methodology becomes critical. This paper represents an exercise in generating a stop list for general text based on the domain dependent and domain independent text collections.

Methodology

Stop lists can be domain independent or domain dependent. They can be created using syntactic classes, or using corpus statistics, resulting in a domain dependent approach for well-defined fields (Abu El- Khair, 2006). This study explore a combination of four techniques of syntactic classes, domain dependent, corpus statistics and expert judgments to obtain the benefits of all approaches. Firstly we reviewed the characteristics of Farsi language and the word structure to identify area of problems for crating stop list. In the next stage a number of 63 articles extracted from 12 different high ranking Persian journal titles of Psychology, Education, and Library and Information Science randomly. To collect the data the articles were typed in a Word format. We manually edited the texts and several fatal spelling errors in the texts were corrected. Then the words automatically segmented. A post-segmentation script was used to join the detached morpheme to the stem and separate the adjoint words prior to word analysis. Syntactic function of the obtained words were examined by using *Loghatnameh Dehkhoda*, *Farhang Moin*, *Farhang Sokhan*, the comprehensive dictionaries in Farsi language.

The research also used corpus created by Hamshahri online Persian newspaper in Iran. The corpus contains 190,206 articles covering a range of different topics such as politics, city news, economics, reports, editorials, literature, sciences, sociology, foreign news, sports, etc. The size of the documents varies from short news to rather long articles with the average of 1.8 KB. More important it consists of real text in everyday use of Farsi that implies it has the sampling and representative property (Darrudi *et al.*, 2004). The data collected through above mentioned techniques function as basis for data analysis. During the stages the research also called the expert judgments.

Characteristics of the Farsi language and area of difficulties

The effectiveness and ability of any IR system to perform efficiently in a language other than English depend on the capability of the system to conform to specific language in use, and understanding the characteristics of the language under study is of great importance (Abu El-Khair, 2006). This section presents an overview of the major characteristics of Farsi language and the main challenges encountered in the producing a stop word list for Farsi.

Farsi, also known as Persian, is an Indo-European language. It is the official language of Iran and is spoken in Afghanistan, Tajikistan and parts of Uzbakistan. A distinguishing characteristic of the Farsi language is the right to left orientation in writing. Farsi uses an expanded form of Arabic alphabet, introducing letters like ch, p, zh, and g. The Farsi alphabet consists of 32 characters that are consonants and vowels: ا، ب،پ،ت،ث،ج،چ،ح،خ،د،ذ،ر ،ز ،ژ ،س،ش،ص،ض،ط،ظ، ع، غ،ف،ق،ك،گ،ل،م،ن،و ،ه،ي! Moreover, there is a symbol (ء) that can be added to the character or the word and change the pronunciation of the character or the word, which makes a total of thirty-three characters. Letters are distinguished by one (ten cases), two (three cases),or three dots (five cases) placed above or below the letter. Three long vowels, AA, EE, and OO are also represented by letters. Short vowels for A, E and O have no letters and may be shown by diacritic marks. In addition, diacritic mark احتمالاً is used as final overbar stroke to make an adverb form such as احتمالاً (probably). In some cases short vowels can change the meaning of a word bases on their position, e.g. "پُر"(full), "پَر (feather). Some of these usages, however, are limited to colloquial speech and they are rarely used in written text, which results in different possibilities of word analysis.

Farsi plural are formed more regularly by addition of suffixes: "ها"،"ان and words borrowed from Arabic "ات". The plural might be formed irregularly depending the root and singular form of the word and a complete reformulation of the word (سبب، (cause), (means, equipments) اسباب).

Every letter in Farsi is pronounced as a word cannot be used to represent one character except the waw "و" which means "and. This letter comes in its separate format in a large portion of words in Farsi. Acronyms or abbreviations are not common in Farsi; there are some abbreviations in Farsi that are presented with one-letter words. For example, (ص) stand for صلي الله عليه و آله (praise and greeting to God/Mohammad and his descendants) in written Farsi; therefore, except for "و" (and), all one letter words are abbreviations. Farsi is written using the Arabic script, each character may have up to four different presentation forms in writing according to its location in the word or syntactic situation: initial, middle, final and isolated. For example, there are four forms for "غ"; an initial "غار" (cave), a medial "مغول" (Moghul), a final "تيغ" (razor) and an isolated "باغ" (garden). Therefore, each letter may have up to four different shape based on connectivity and its occurrence at the word. The initial form indicates that no element is attached to the element from the right; there is no attaching character before it, but there is one followed the character. Characters are in medial form if they have an attached character both before and after them. The final form denotes that the character is at the end of a word. The final forms can therefore be used to mark word boundaries. The isolated form is a standalone character (Taghva *et al.*, 2003a, b). In Farsi, words may be written as connected or separated characters "ميروم" ، "ميرِوم" (I am going). Thus, a single word may have got different presentation forms. In addition, some letters do not connect to others at all, "زود" (soon, quickly, early).

In Farsi texts, word boundaries are identified by using space as word delimiters. But, the inconsistent usage of the whitespace in the texts gives rise to problems in detecting word and boundaries. The optional nature of whitespace causes distinct words to appear as single token(e. g. بیان‌کردن = to express), also raises issues in the detached morphemes such as رفته اي (you have gone). The inflectional morphemes such as "ترین"،"تر"،"ها"،"مي"* can appear either as bound to the host, as free affixes separated by final form character, or separated with an intervening space (e.g. کتابها، کتاب‌ها، کتاب ها = books). The discontinuity may occur in the word structure such as نمایه ساز (indexer). There exist lexical elements, such as the preposition به, the determiner "این", the postpositions "را", or the relativizer "که" that usually appear a separate words in written text, but may appear attached to the adjacent word (e.g. بشدت = intensively). Similarly, a number of pronominal or verbal clitic elements may occur on various of speech categories (e.g. روبروست = is facing). The element within a noun phrase are linked by the enclitic particle called ezafe. This morpheme is usually an unwritten vowel (Megerdoomian, 2004). Lack of an ezafe morpheme delimit the boundary of the pronoun and the proper name in written text as علي شريعتي (Ali Shariati). In such cases the word constructs two words. There are also a large number of multi word expression in Farsi. These include phrasal verbs, compound nouns and lexical units (unit-like element) such as بنابراین (therefore), براساس (based on). The presence of these multi words make the segmentation difficult because of optional whitespace within them (Megerdoomian, 2004). Farsi verbs are modified more extensively than English verbs. Farsi verbs vary form according to tense, person, negation, and mood. Therefore, a given verb may have a score of variation (Taghva *et al.*, 2003a, b). For instance, Figure 1 shows all variations (present participle) of the verb رفتن (to go). According to the figure it has 136 variations.

All these complexities together and multi meaning and multi functioning of the words lead to difficulties in recognizing word boundaries and producing invalid words in the word extraction stage and make processing and identification of stop words in Farsi language more difficult than it appears. Reviewing of the Persian linguistic and grammar literature (Safavi, 1981; Meshkatoddini, 2005; Bateni, 2003) revealed that words in Farsi such as other languages have two distinct level of representation: semantic and syntactic representation. Syntactic words have a finite state and semantic ones have an infinite state. As is mentioned in the related literature stop words serve only a syntactic function. Instead, they are used just because of grammar and carry no significant information (Zou *et al.*, 2006). Therefore, possible words that may be considered as stop words should be collected from the different syntactic classes in Farsi in a systematic way to ensure the completeness of the list. From the viewpoint of linguistics, Farsi stop word usually will be those words with the following word categories: adverbs, pronouns, prepositions, determiners, conjunctions, interjection, ordinal numbers, lexicon units, auxiliaries and light verbs.

As Megerdoomian (2004) believes, listing of the unit-like elements (e.g. بنابراین = therefore) can improve the accuracy of Farsi segmentation significantly. Moreover the subparts of these lexicon units may not appear as index or search terms in Farsi information retrieval system. Most verbal constructions in Farsi are formed using a light verb such as کردن (to do), (to hit), زدن(to give) دادن The number of these verbs is limited but their constructions are extremely productive in Farsi. Light verbs in Farsi behave like single lexical verbs and also function as verbs, when they combine

ماضي ساده: رفتم، رفتي، رفت، رفتيم، رفتيد، رفتند(نرفتم، نرفتي، نرفت، نرفتيم، نرفتيد، نرفتند

ماضي استمراري:مي‌رفتم، مي‌رفتي، مي‌رفت، مي‌رفتيم، مي‌رفتيد، مي‌رفتند(نمي‌رفتم

ماضي نقلي: رفته‌ام، رفته‌اي، رفته‌است، رفته‌ايم، رفته‌ايد، رفته‌اند(نرفته‌ام

ماضي بعيد: رفته بودم، رفته بودي، رفته بود، رفته بوديم، رفته بوديد، رفته بودند(نرفته بودم

ماضي التزامي: رفته باشم، رفته باشي، رفته باشد، رفته باشيم، رفته باشيد، رفته باشند(نرفته باشم

ماضي مستمر: داشتم مي‌رفتم، داشتي مي‌رفتي، داشت مي‌رفت، داشتيم مي‌رفتيم، داشتيد مي‌رفتيد، داشتند مي‌رفتند(نمي‌رفتم

مضارع ساده: روم، روي، رود، رويم، رويد، روند(نروم

مضارع اخباري: مي‌روم، مي‌روي، مي‌رود، مي‌رويم، مي رويد، مي‌روند(نمي‌روم

مضارع التزامي: بروم، بروي، برود، برويم، برويد، بروند(نروم

مضارع مستمر:دارم مي‌روم، داري مي‌روي، دارد مي‌رود، داريم مي‌رويم، داريد مي‌رويد، دارند مي‌روند(نمي‌روم

آينده: خواهم رفت، خواهي رفت، خواهد رفت، خواهيم رفت، خواهيد رفت، خواهند رفت(تخواهم رفت

امر: دوم شخص مفرد: برو (نرو

دوم شخص جمع: برويد (نرويد

Figure 1.
Variation of the verb رفتن

with different non-verbal items (Karami-Doostan, 1997). Farsi light verbs have many variations and each variation is a poor index or search term. However, a large number of these variations because of their infrequent occur failed to make our stop list. Personals pronouns can appear either as free or as clitics. Although these cliticized pronouns have the same surface form, they can have different functions depending on the part of speech or syntactic context that they appear on. In Farsi just free forms of pronouns can be collected as stop words. The use of a stemming algorithm may filter out attached words such as cliticized pronouns.

Domain-based

Text processing of 63 articles resulted in a total of 248,552 occurrences of words. Assigning grammatical categories to words in a text is an important and difficult

component of NLP. The syntactic form of these words were identified based on the mentioned dictionaries, personal and expert judgments. The frequency of the words were also calculated. Then the stop list was constructed on the syntactic function and frequency of the words. Since rule-based technique can accurately annotate grammatical function of the words, it is used to analyze and mark up a given texts. Statistical technique (frequently) is then applied to disambiguate the result or to guess the stop words for the list. The resulting list constructed of 1,291 words in the three domains. Further analysis showed that stop words represents about 39 per cent and symbols 7.4 per cent (both 46.47 per cent) of the tokens in the collected texts. The punctuation marks that appeared in the body of the 63 articles are: – _ #) (*} {!. . . /:] [„ ? “ “ ،.

Furthermore, the chi-square test was used to judge whether word distribution differences between the three domains can be considered statistically significant or not. The differences was among 20 per cent of the infrequent words ($p = 0.05$). The distribution of word frequency showed that the highest frequent stop words have a quite stable distribution in different documents and domains.

We wanted a large list to keep many useless words out of indexes, but browsing the data by a small group of experts it was found that many important words as potential index or search terms were occurred in the list; also the presence of such a words may have a negative effect on recall of Farsi IR system. Based on the expert recommendation these words (545 words) were removed, and a list of 746 words produced at this stage. The top ten words are: و ، در ، به، که، از ، این، را، است، با، براي (and, at, to, that/which/who, of/from, this, را = particle serving as a sign of the direct object, is, by/with, for)

Corpus-based

In the next stage, Hamshahri corpus statistic were used to create a corpus-based stop list words occurring more than 20,000 times. Preliminary examination of the corpus word frequency statistics and resulting stop list suggested that this was a reasonable heuristic. A total of 758 words and symbols (737 + 21) met this criterion. The top ten words are: براي، و ، کشور ، در ، ایران، به، از ، کنند، دارد، تهران (for, and, country, at, Iran, to, of/from, they do, have, Tehran). Analysis of the punctuation marks appeared within the structure of the corpus revealed that these symbols are the same that were illustrated in the previous pages. Browsing the data from the created list it was also obvious that many words including invalid words, different presentation forms as well as words important as index or search terms such as country, Iran, and Tehran among the top words. A proper segmentation of Farsi texts is required before construction of, because the word boundaries are not clear in Farsi texts. Therefore, a manual check was done to remove any content bearing words, which may not be considered conventional stop words, from this list. The removal of these words (315 words) in another decision based on personal judgment, and syntactic rules (word categories) were identified in this study. The resulting list contained 422 words, as second list of stop words. Comparing the two word list indicated that the content of these stop word list vary in a lot from each others, but quite stable for 246 words (Table I). The chi-square test result ($P = 0.01$) showed that the difference between the two techniques distribution is significant at the 99 per cent level of confidence. This result indicates that for producing a complete and reliable stop word list in Farsi both statistical and

rule-based as well as domain dependent and independent methodologies should be combined.

Aggregation

The two list were aggregated together to generate the final one. A combination of these two observations redefined the stop words as those words with stable and high frequency in documents. This list generated according to the above mentioned techniques revealed the features of stop words in different manners, which are all quite reasonable. After removing 246 words in common between the two lists, the stop list is brought to its final of 922 words, that should be maximally efficient and effective in filtering the frequently occurring and semantically neutral words in Farsi texts. These words are listed in alphabetical order in the Appendix (Figure A1). It contains verbal and nonverbal stop words in Farsi almost none of which would be potentially good index or search terms. Our experiments are based on Hamshahri corpus and a text collection in psychology, education, and library and information sciences. It is believed that those stop lists produced for a language in general will not be able to function appropriately in specific domains. Therefore, we recommend that stop lists can be prepared for each subject area separately, i.e. for biology, chemistry, physics, etc. However, the lack of common large collection has been a major draw back in Farsi text retrieval experiments.

English, Arabic and Farsi stop word list comparison

The final aggregated list was checked against the stop word list of SMART (Lewis *et al.*, 2004), which is a standard and well known in English language. We find that about 43.4 per cent (i.e. 248 words out of 571 words) of the English stop words have corresponding words in Farsi stop word list. For instance, "و" (and), "در" (at), "به" (to), "که" (that/which/who), "از" (of, from), "این" (this), "است" (is), "با" (by, with), "برای" (for). The large number of stop words is due to the characteristic of the Farsi language. However, the specialty of Farsi stop word list is the presence of lexical units, lack of clitic personal pronouns and also that some words might have the same meaning, like "of", "from",both of which means "از".

Because of the close relationship between Arabic and Farsi and the presence of a lot of Arabic loan words in Farsi, we also compared the overlap of our Farsi stop word list with a general Arabic stop word list created by Abu El- Khair (2006). The result showed that only 48 (5.20 per cent) words such as احیاناً، اخیراً، اکثر، الا، اقل are common between the two lists.

Techniques	Number of initial words	Number of removed words	Number of stop words	Number of common words	Number of non-common words
Domain-based	1,291	545	746	246	500
Corpus-based	737	315	422		176
					676
Total	2,028	855	1,173	246	676 + 246 = 922

Table I. Comparison of the domain-based and corpus-based stop word distribution

Adding words for free

Add words to the list according to the following criteria:

- Our list includes the highest occurrence light verbs, each with infinitive and past tense forms. Though a given verbal root may occur frequently, most of its variations occur infrequently. Therefore, add different variation of the light verbs in list.
- Ordinal numbers in Farsi are made with cardinal numbers plus م/om, مي/mi, مين/omin. Add these for that differ from a word already in list.
- Add any single letter that differ from a letter already in list.
- Add any different spelling of the words included in list (به موقع، بموقع = in-time/opportune).
- If a word in the list can take the suffixes ها، ان، ات، تر، ترين, then add the suffixed words (e.g. جديتر = more serious).
- All the suffixed words and unit-like element in the list are written without intervening space, add different presentation forms of these words.
- If a word in the list can take clitics pronouns "ام"، "ات"، "اش"، "مان"، "تان"، "شان", then add the suffixed words (e.g. برايش = for him).

Conclusion

Traditionally stop words are supposed to have included only the most frequent occurring words. Browsing the data from our lists, domain-based and corpus-based, revealed that stop lists have tended to include infrequently occurring words, and have not included many frequently occurring words as also identified by Fox (1992). Therefore, it can be concluded that statistical model may have poor performance in identifying stop list in Farsi. Since automatic word processing is based purely on spelling and space of words, because of different spelling and space character may or may not be occurred, certain ambiguities arise in a computational segmentation and processing of Farsi texts. In fact the same surface form can represent different morphemes. So our Farsi stop word list facilitates the process of the word segmentation in Farsi information retrieval by increasing the accuracy of segmentation. The stop list that we have generated can serve as the basis for stop lists for specialized databases, or as a list for general Farsi literature.

The major conclusions that can be drawn from this exploratory work are:

- Multi analysis method would have a good performance in Farsi language processing.
- All the words under different syntactic classes, except letters, may not be considered as conventional stop words, as some of them are words important as index or search terms.
- The extremely common words are quite stable among different disciplines; the differences are among the infrequent words.
- Stop words represent approximately 39 per cent of the tokens in Farsi texts, therefore, excluding stop words, especially in full text search system, save space about 39 per cent and speed up searches with little effect on the quality of results.

- About 43 per cent of the English stop words have corresponding words in Farsi stop word list, it can be concluded that translating an English stop word list even augmenting it with high frequency words from the corpus may not be suitable for Farsi texts.

References

Abu El- Khair, I. (2006), "Effect of stop words elimination for Arabic information retrieval: a comparative study", *International Journal of Computing & Information Sciences*, Vol. 4 No. 3, pp. 119-33.

Bateni, M.R. (2003), توصيف ساختاري دستوري زبان فارسي بر بنياد يك نظريه عمومي زبان, Amir Kabir, Tehran.

Darrudi, E., Hejazi, M.R. and Oroumchian, F. (2004), "Assessment of a modern Farsi corpus", *Proceedings of the 2nd Workshop on Information Technology and its Disciplines (WITID) 2004, ITRC, Kish Island, Iran.*

Falahati Qadimi Fumani, M.R. and Ramachandra, C.S. (2008), "The concept of stop words in Persian chemistry articles: a discussion in automatic indexing", available at: http://bibliotecavirtualut.suagm.edu/Glossa2/journal/dec2008

Fox, C. (1990), *Lexical Analysis and Stop List*, Prentice-Hall, Upper Saddle River, NJ.

Fox, C. (1992), "A stop list for general text", *SIGIR Forum*, Vol. 24 Nos 1/2, pp. 19-35.

Francis, W. and Kucera, H. (1982), *Frequency Analysis of English Usage*, Houghton Mifflin, New York, NY.

Karami-Doostan, M.R. (1997), "Light verb construction in Persian", PhD dissertation, University of Essex, Colchester.

Korfhage, R.R. (1997), *Information Storage and Retrieval*, John Wiley, New York, NY.

Lewis, D., Yang, Y., Rose, T. and Li, F. (2004), "Rcv1: a new benchmark collection for text categorization research", *Journal of Machine Learning Research*, Vol. 5, pp. 361-97.

Megerdoomian, K. (2004), "Developing a Persian part of speech tagger", *Proceedings of the 1st Workshop on Persian Language and Computers, Tehran University, Tehran, May.*

Meshkatoddini, M. (2005), دستور زبان فارسي: واژگان و پيوندهاي ساختي, SAMT, Tehran.

Safavi, K. (1981), درآمدي بر زبانشناسي, Bongan Tarjome va Nashr, Tehran.

Schauble, P. (1997), *Multimedia Information Retrieval: Content-based Information Retrieval from Large Text and Audio Databases*, Kluwer Academic Publishers, Boston, MA.

Taghva, K., Beckley, R. and Sadeh, M. (2003a), "A list of Farsi stop words", Technical report 2003-01, Information Science Research Institute, University of Nevada, Las Vegas, NV.

Taghva, K., Young, R., Coombs, J., Beckley, R., Sadeh, M. and Pereda, R. (2003b), "Farsi searching and display technologies", Technical report NV 89154-4021, Information Science Research Institute, University of Nevada, Las Vegas, NV.

Van Rijisbergen, C.J. (1975), *Information Retrieval*, Butterworths, London.

Zou, F., Wang, F.L., Deng, X. and Han, S. (2006), "Automatic identification of Chinese stop words: advances in natural language processing", *Research in Computer Science*, Vol. 18, pp. 151-62.

Further reading

Megerdoomian, K. (2000), "Unification-based Persian morphology", *Proceedings of CICLing 2000, Centro de investigacion en computation-IPN, Mexico* (edited by A. Gelbukh).

Appendix

اکثر	ازاین‌روي	آیا	آخ
اکثرا	ازاین‌قرار	اتفاقاً	آخر
اکثریت	ازاین‌گذشته	اتفاقي	آرام آرام
اکنون	ازجمله	احتمالاً	آشکارا
اگر	ازچه‌قرار	احیاناً	آمد
اگرچه	ازروي	اختصارا	آمدن
اگرنه	ازقبیل	اخیر	آن
الآن	ازلحاظ	اخیرا	آنان
الا	ازنظر	از	آناني
البته	ازهمین‌رو	ازآن	آنجا
الي	اساساً	ازآن زمان به بعد	آنچنان
اما	است	ازآن‌به‌بعد	آنچنان‌که
امروز	اشتباهاً	ازآن‌پس	آنچه
امروزه	اصطلاحاً	ازآنجا که	آنچه‌که
امسال	اصلا	ازآنجاییکه	آنرا
امشب	اصولاً	ازآن‌جهت	آنقدر
امور	اضافه‌براین	ازآن‌جهت‌که	آنکس
انجام	اظهار	ازآن‌رو	آنکه
اندك	اعلام	ازاین	آنگاه
اندکي	اغلب	ازاین به بعد	آنگونه
انگار	افزون‌براین	ازاین‌پس	آن‌وقت
او	افسوس	ازاین‌جهت	آنها
اول	اقل	ازاین‌جهت‌که	آنهایي
اولا	اقلاً	ازاین‌دست	آورد
اي	اکتساباً	ازاین‌رو	آوردن
به‌جرأت	بعداز	بایستي	ایشان
به‌جز	بعدازآن	بتدریج	این
به‌جهت	بعدازاین	بجز	اینان
به‌خاطر	بعدازاینکه	بخش	اینجا
به‌خاطراینکه	بعدازظهر	بخشي	اینچنین
به‌خصوص	بعدها	بخصوص	این‌طور
به‌خلاف	بعضاً	بخوبي	این‌قدر
به‌خوبي	بعضي	بدان	اینك
به‌خود	بعضي دیگر	بدان جهت	اینکه
به‌خودي‌خود	بلادرنگ	بدانجا	این‌گونه
به‌درستي	بلافاصله	بدانها	اینها
به‌دقت	بلکه	بدون	این‌همه
به‌دلخواه	بله	بدون‌آنکه	با
به‌دلیل‌آن‌که	بلي	بدون‌اینکه	باآنکه
به‌دلیل‌اینکه	بنا به	بدین	بااطمینان

Figure A1.
List of Farsi stop words

(*continued*)

بمراحتي	بنابر	بدين‌ترتيب	بااين‌حال
بمراستي	بنابراين	بدينجا	بااينكه
بمرغم	بود	بدينسان	بااين‌وجود
بمروز	بودن	بدين‌قرار	بااين‌همه
بمروشني	به	بدين‌معني‌كه	باتوجه‌به‌اينكه
بمزعم	به‌آساني	بر	بارها
بمزودي	به‌اجمال	بر روي	بازهم
بمسادگي	به‌استثناي	برابر	باعلاقه
بمسبب	به‌اضافه	براثر	بالا
بمسبب‌آنكه	به‌اين ترتيب	براساس	بالاخره
بمسبب‌اينكه	به‌بالا	براي	بالاخص
بمسختي	به‌بعد	براي‌آن‌كه	بالاي
بمسرعت	به‌تازگي	براي‌اينكه	بالضروره
بمسوي	به‌تدريج	برحسب	بالطبع
بمسهولت	بهتر	برخلاف	بالعكس
بمشدت	به‌ترتيب	برخي	بالغ بر
بمشرط‌آن‌كه	به‌تفكيك	برطبق	بالنتيجه
بمشرط‌اينكه	به تمامي	برعكس	باوجود
بمصورت	به‌تنهايي	بس	باوجودآنكه
بمصورتيكه	به‌توسط	بسا	باوجوداين
بمطور	به‌جا	بسي	باوجوداينكه
بمطوركلي	به‌جاي	بسيار	باوجودي‌كه
بمطوري‌كه	به‌جاي‌آنكه	بسياري	باهم
بمظاهر	به‌جاي‌اينكه	بعد	بايد
به‌عكس	به‌جد	بعدا	بايست
ثانيا	پيش‌ازاين	به‌هيچ‌وجه	به‌علاوه
جا	پيشتر	بي	به‌عنوان
جاي جاي	پيش‌روي	بي‌آنكه	به‌غيراز
جدا	پيشين	بي‌پايان	به‌قدر
جدا	پي‌گير	بي‌ترديد	به‌قدري
جدا از	تا	بي‌توجه	به‌قرار
جدا از هم	تاآنجاكه	بي‌جا	به‌قول
جداگانه	تاآنكه	بي‌چون و چرا	به‌كرات
جداي از	تااندازه‌اي	بيرون	به‌كل
جدي	تابه‌حال	بيش	به‌كلي
جديد	تاجاييكه	بيش‌از	به‌كندي
جديدا	تاحدودي	بيش‌ازآنكه	به‌گرمي
جز	تاحدي	بيش‌ازاينكه	به‌لحاظ
جزء به جزء	تاكنون	بيشتر	به‌لحاظ اينكه
جلو	تام و تمام	بي‌شك	به‌مثابه

(*continued*)

Figure A1.

جلوتر	تاوقتي‌كه	بي‌گمان	به‌مجرداينكه
جلوي	تحت	بي‌مناسبت	به‌محض‌آنكه
جمعاً	تدريجاً	بين	به‌مراتب
جهت	ترجيحاً	بينابين	به‌منزله
چرا	تصريح	بي‌نتيجه	به‌منظور
چراكه	تصريحاً	بي‌وقفه	به‌منظوراينكه
چقدر	تعدادي	پارسال	به‌موجب
چگونه	تعمداً	پاره‌اي	به‌موقع
چنان	تقريباً	پايين	به‌ناچار
چنانچه	تك تك	پريروز	به‌ندرت
چنانكه	تلويحاً	پس	به‌نسبت
چند	تمام	پس و پيش	به‌نظر
چندان	تمامي	پس‌از	به‌نوبه‌خود
چندمين	تمام وقت	پس‌ازآنكه	به‌نوعي
چندي	تماماً	پس‌ازآن	به‌واسطه
چندين	تند تند	پس‌ازاين	به‌واسطه‌اينكه
چندين‌بار	تنگاتنگ	پس‌ازاينكه	به‌واقع
چنين	تنها	پشت	به‌وسيله
چو	تو	پي	به‌وضوح
چون	تواماً	پي‌درپي	به‌ويژه
چه	توانست	پيرامون	به‌هرحال
چه‌آنكه	توانستن	پيش	به‌هرروي
چهارم	توسط	پيشاپيش	به‌هرصورت
چه‌بسا	ثالثاً	پيش‌از	به‌هم
چه‌چيز	ثاني	پيش‌ازآنكه	به‌هيچ‌روي
رو به گسترش	درمورد	خير	چه‌چيزهايي
روبرو	درميان	خيلي	چه‌چيزي
روز به روز	درنتيجه	دائم	چهره به چهره
روزانه	درنهايت	دائماً	چه‌طور
روزمره	درواقع	داخل	چيز
روي	درون	داد	چيزي
روي هم	درهرحال	دادن	چيست
روي هم رفته	درهرصورت	دارد	حاشا و كلا
الزاماً	دريغ	داشت	حاكي
زماني	دريغا	داشتن	حال
زمانيكه	دسته‌دسته	در	حال‌آنكه
زود	دقيقاً	در ازاي	حالا
زهي	دگرباره	دراثر	حال‌كه
زياد	دگرگون	دراين‌باره	حتماً
زير	دوباره	دراين‌ميان	حتي

Figure A1.

(continued)

زيرا	دوتا دوتا	درباب	حتي‌المقدور
زيرا كه	دوراز	درباره	حداقل
س	دوم	درباره ي	حداكثر
سابقاً	دومي	دربرابر	حدود
سالانه	دهم	دربه‌در	حرف به حرف
ساليانه	دير	درپي	حسب
ساير	ديروز	درتحت	حقيقتاً
سپس	ديشب	درثاني	حكماً
سراسر	ديگر	درجهت	حول
سرانجام	ديگران	درحالي كه	خارج‌از
سريع	ديگري	درحاليكه	خاصه
سريعاً	ذاتاً	درحقيقت	خامساً
سوم	ذيلاً	درحين	خداحافظ
سومي	راجع‌به	درخصوص	خصوصاً
سوي	راحت	درست	خواست
سهل‌الوصول	راساً	درصورتي	خواستن
سهواً	راست	درصورتي‌كه	خواه
شايد	راستا	درطول	خواه‌ناخواه
شبانه	راستي	درطي	خود
شبانه‌روز	رسماً	درعين	خودبه‌خود
شتابان	رفته رفته	درعين‌حال	خودبه‌خودي
شتاب‌گونه	رو به افزايش	دركل	خوش
شخصاً	رو به تزايد	دركنار	خوشبختانه
شد	رو به زوال	درمجموع	خويش
شدن	رو به سستي	درمقابل	خويشتن
گرفت	قريب به اتفاق	عمدتاً	شديداً
گرفتن	قطعاً	عمده	شما
گفت	قوياً	عملاً	شماري
گفتن	كاش	عموم	ص
گوناگون	كاشكي	عموماً	صادقانه
گويا	كافي	عميقاً	صد البته
گويي	كامل	عنقريب	صراحتاً
گهگاه	كاملا	عيناً	صرفاً
لا	كجا	غالباً	صريحاً
لااقل	كدام	غير	صفحه به صفحه
لاجرم	كدامند	غيراز	صميمانه
لذا	كدام‌ها	غيرازاين	ضرورتاً
لزوماً	كدام‌يك	غيرممكن	ضمناً
لطفاً	كرد	فاقد	طبعاً
ليكن	كردن	فيها	طبيعتاً

(*continued*)

Figure A1.

مؤخر	كساني	فراتر	طرف
مآلا	كسي	فردا	طريق
ما	كلا	فعالانه	طور
ما قبل	كم	فعلا	طي
مادام كه	كم كم	فقط	ظاهرا
مادامي	كم و بيش	فلان	ع
مادامي‌كه	كمااينكه	فلذا	عاجزانه
مانند	كمابيش	فورا	عاري از
ماهانه	كماكان	فوري	عاقبت
ماهيتاً	كمتر	فوق	عبارتند
مبادا	كمي	فوق‌العاده	عجب
متأسفانه	كنار	في‌الواقع	عجولانه
متعاقباً	كنون	في‌نفسه	عقب
متفاوت	كنوني	ق.م.	علاوه بر
متقابلا	كه	قابل	علاوه‌برآن
مثال	كي	قاطعانه	علاوه‌برآنكه
مثل	كيست	قاعدتاً	علاوه‌براين
مثلاً	گاه	قبل	علاوه‌براينكه
مجدد	گاهي	قبلا	علناً
مجدداً	گذرا	قبل‌از	علي‌الاصول
مجموعاً	گذشته از	قبل‌ازآن	علي‌الظاهر
محتاطانه	گذشته ازآنكه	قبل‌ازاينكه	علي‌رغم
محكم	گذشته‌ازاين	قدر	عليرغم اينكه
مختصراً	گرچه	قدرمسلم	عليه
مختلف	گردد	قدري	عمداً
همان‌طوركه	وقتي	موارد	مخصوصاً
همان‌طوري‌كه	وقتي‌كه	موجود	مدام
همان‌قدر	وگر	مورد	مداوم
همان‌گونه‌كه	وگرنه	موقعي‌كه	مدت
همانند	ولاغير	ميان	مدتي
همانى	ولي	مي‌بايست	مرا
همچنان	وليكن	ناچار	مرتب
همچنان‌كه	وي	ناگزير	مرتباً
همچنين	ويژه	ناگهان	مرحله‌به‌مرحله
همچون	ه.ش.	نبايد	مستقلا
همديگر	ه.ق.	نتيجتاً	مستقيماً
همزمان	هان	نخست	مستمراً
همسو	هر	نخستين	مسلماً
همگام	هرآنچه	ندرتاً	مشتركاً
همگان	هرازچند گاهي	نزد	مشخص

Figure A1.

(*continued*)

همگي	هرازگاهي	نزديك	مشروط برآنكه
همواره	هرجا	نزديكتر	مضاف بر
همه	هرچقدر	نسبتاً	مطلقا
هميشگي	هرچند	نسبت‌به	مطمئناً
هميشه	هرچندكه	نشان	مع الاسف
همين	هرچه	نظراً	مع ذالك
همين‌طور	هرساله	نظربه‌اينكه	معدود
همين‌كه	هركدام	نظير	معمولاً
هنگامي‌كه	هركس	نگاه	معمولي
هنوز	هركسي	نوعاً	مغرضانه
هيچ	هركه	نوعي	مقابل
هيچ‌چيز	هرگاه	نه	مقدار
هيچ‌كدام	هرگز	نه‌آنكه	مكرر
هيچ‌كس	هرگونه	نهان	مكرراً
هيچ‌گاه	هروقت	نهايتاً	مگر
هيچ‌گونه	هريك	نه‌اينكه	مگرآنكه
هيچ‌يك	هزاران	نه‌تنها	مگراينكه
يا	هزارها	نه‌چندان	ملاحظه
ياآنكه	هست	نه‌فقط	ملزم به
ياينكه	هفتگي	نيز	ممكن
يافت	هم	و	من
يافتن	هم اكنون	واقعاً	منتهي
يعني	هم اينك	واقعي	منحصراً
يقيناً	همان	واي	منحصربه‌فرد
يك	همانا	وراي	منظور
يك‌كمي	يكديگر	يكجانبه	يكايك
يك‌لحظه	يكزمان	يك‌جور	يكبار
يكنواخت	يكسال	يك‌جوري	يكباره
يكي	يكسره	يك‌جهت	يك‌به‌يك
	يك‌سري	يك‌چند	يكپارچه
	يك‌كم	يك‌دم	يكجا

Figure A1.