Persian Basic Words based on Newspaper Texts

Reza Morad Sahraei[1]
Amirhossein Mojiri Foroushani[2]
Morvarid Talebi[3]

Received: 11/06/2018
Accepted: 01/01/2019

Abstract

No information can be transmitted without familiarity with language words. Teaching vocabulary is one of the most important components of foreign language teaching that can affect all the four main language skills (listening, speaking, reading and writing). The first step in teaching vocabulary is to access the list of basic words. According to the studies conducted in the related area, high-frequency words as well as the basic vocabularies are significant in language teaching since they are easy to learn and frequently used in everyday conversations. Frequently-used words list or frequency dictionary is a set of words that have more repetition in a collection of texts (corpus). Basic words are generally extracted from the architecture of the corpus-based researches, and the output of each linguistic corpus can be a basic words list (depending on the language and type of the corpus texts).

From 1897 there are many lists of basic words in different languages of the world. English language researches is more than any other researches in this field. Since the year 1971, researches have also been conducted in Persian to extract frequent words.

Thorndike (1921) presented 30,000 basic English words. In 1923, Ogden and Richards listed 850 basic words of English. Dolch (1936) listed 220 and West (1953) presented 2,000 basic English words. Coxhead (2000) has derived basic words in four areas of art, commerce, law and science. In 2001, Verlinde and Selva provided a list of frequent words in French. Also, 100 and 1000 English basic words were extracted by Fry et al. (2000). Jones and Tschirner (2006) extracted 4,307 frequent German words. Davies and Gardner (2010) extracted 1,000 to 5,000 basic English words. The list of 100 frequent words of Oxford English dictionary and the 3,000 basic words of Langman's dictionary are other instances.

In Iran, Barahani (1975), Imen (1978), Safarpour (1991), Tahriryan (1994), Hasani (2005), Gharavi Qouchani (2006), Doroody et al. (2008), Alahmad et al.

[1] PhD in Linguistics, Associate Professor at Department of Linguistics, Allameh Tabataba'i University, (Corresponding Author); sahraei@atu.ac.ir
[2] MA in Teaching Persian to Foreign Language Learners, Researcher at Sa'di Foundation; amojiry@saadifoundation.ir
[3] MA in Teaching Persian to Foreign Language Learners, Allameh Tabataba'i University; morvarid_talebi@atu.ac.ir

(2009), Bijankhan (2011), Nematzadeh et al. (2011) were among the scholars of studying basic words.

This research has two main stages: a) extracting texts and registering in database: 8 persons within 100 working days, each day extracted 3 passages with an average of 500 words from three newspapers in one of the seven different areas (including culture, society, politics, sports, fiction, economics and science), resulting in a corpus with 1,203,589 words (2401 texts).

The software used for this project was written specifically for this research using the PHP programming language and is a web-based software. Types of words in the software are “name”, “verb” (and in particular “compound verb”), “preposition”, “proper noun”, “adjective” and “adverb”. Also, in this corpus, each text has metadata of “type of text” (cultural, social, political, etc.), the name of the newspaper, the date of printing, the date of the text typing, the date of frequency extraction and the name of the researcher. A list of the collections was also made to identify broken plurals. For example, "آثار: اثر". Also, a list of inflectional affix (prefixes and suffixes) and rules governing them was provided. This list specifies what suffixes or prefixes any type of word can take. For example, verb can start with the "می" prefix, the "ات"suffix can be added to the word "آیه" but before that, the letter "ه" should be deleted from the end of the word.

Each word in the software also has attributes like prefix, suffix, word root, word category, text numbers, word frequency, and main word (the word used in the text).

b) Labeling Words: Labeling involves specifying the word category (noun, adjective, verb or preposition), and the lemma. Words with derivational affixes and without inflectional affixes and clitics were recorded. The verbs were recorded as infinitive. Compound verbs were recorded in an infinitive form, and their nominal and verbal parts were not separated.

After the end of the previous stage, the words of the texts (including 1,203,598 words) were obtained along lemma of each word. After corrections such as label correction and lemma correction, the “Basic words” table (including 2,150 words with a frequency of 50 and above) was obtained. In addition, the list of 50 most frequent names, prepositions, adjectives and 20 most frequent adverbs in Persian were also obtained. Then, the specific base words of each topic (including 500 specific words of each topic with a frequency of 12 and above) were obtained.

Keywords: corpus-based research, newspaper texts corpus, teaching vocabulary, basic words, high-frequency Persian words