Chapter III

STRUCTURE OF LEXICOGRAPHY
3. STRUCTURE of LEXICOGRAPHY

3.1 Introduction

Landau (1984: 5) writes: "A dictionary is a book that lists words in alphabetical order and describes their meaning." In other words, a dictionary is a reference book of arranged lexical units in alphabetical order and it describes the meaning of the lexical units. Besides "Modern dictionaries often include information about spelling, syllabification, pronunciation, etymology, usage, synonyms, and grammars, and sometimes illustrations as well." Normally the dictionary has two structures that are macro structure and micro structure. The structuring of material of any dictionary must arrange to help the users by means of instructions provided.

Macro structure refers to the list and organization of the lexical items entered in the dictionary, the lemmas or headwords. Macro structure is an overall word list structure that allows the compiler to locate information in a dictionary. ‘The macro structure is an equivalent of the English word list (Bejoint 2000:11). The dictionary entries must be arranged in a particular order and in a layout that is in accordance with the aims and usage of the dictionary.

There are several ways of arranging the words in a dictionary and dictionaries differ greatly in their entry policies. The commonest and most frequent way of arranging entries in dictionary is the alphabetic arrangement. The other way of arranging the headwords are discussed in separate section named as ordered set. Most of the dictionaries deal with the homographs by entering them separately but some other dictionaries deal with the homographs in a single entry with meaning separation.

The internal structure of a dictionary entry, often called as microstructure of the dictionary. It refers to the arrangement of the information within the lexical entry.
(Jackson 2002: 26). The micro structure involves not only the order of appearance of various information categories but also their typographical form such as type face and type size, punctuation and lexicographical symbols etc. Various information categories are arranged in a suitable manner. Comparing to the macro structure, the microstructure of a dictionary is a rigid structure. All modern dictionaries are characterized by the uniformity of the entries, both in terms of their content and their layout (Zgusta 1971: 248). Different dictionaries have different policies for the information provided under microstructure due to the influence on selection of ‘lexically relevant’ information and about presentation of the information in micro structure of a dictionary. The range and type of information given for a lexical entry varies according to the type, size, purpose and user of the dictionary.

Typically, dictionaries have all of the following information under the microstructure of a lexical entry: (a) Spelling, Pronunciation, Transliteration, Tone and intonation, (b) Grammatical categories Inflection and derivation, (c) Etymology, Meaning- definition, Sense or different shades of meaning, Record the changes at the level of form and meaning as well: Synonyms, Antonyms, Super ordinate term or hyponym, (d) Phrases, Supporting or related idioms and proverbs, (e) Citation, Examples, Usage, Semantic domain, Usage level, and (f) Pictorial information. The micro structure customarily provides information on the form, meaning and use of the lemma. Formal information may include spelling and pronunciation, base and inflected forms and syntactic category. The semantic information includes definitions or explanations of literal and figurative, denotative and connotative meanings. These may take the form of synonyms or nearest form of synonyms, paraphrases or formulae. They are usually supplemented by paradigmatic information on lexical fields involving synonyms, antonyms, or hyponyms and by systematic information on lexical
collocation, grammatical colligation and complementation, and on use in idioms, proverbs, and other fixed phrases and chunks; and by pragmatic information marking on register, frequency, currency, style, status, and subject area. They may be complemented by pictorial illustrations, authentic, adapted and constructed textual examples and usage notes.

3.2 Persian Language

The Persian language is categorized as a branch of Indo-European languages and is the official language of Iran, Afghanistan and Tajikistan. It is also spoken in some other countries in the Middle East. It has some characteristics that necessitate usage of different information retrieval algorithms. Morphological analysis of the language is relatively hard because of its grammatical rules. For example the word “خبر” (khabar (news)) is an Arabic word that is used in Persian. This word has two plural forms in Persian “خبرها” (akhbar)” and “خبرها” (khabarha)”, the first plural form follows Arabic grammatical rules and the second plural form is formed by the use of Persian rules (AleAhmad, Kamalloo, Zareh, Rahgozar & Oroumchian, 2009).

Persian uses Arabic like script (alphabets) for writing and which consists of 32 characters that are written continuously from right to left. During its long history, the language has been influenced by many other languages such as Arabic, Turkish and even European languages such as English and French. Today’s Persian contains many words from the above languages and in some cases these words still follow the grammar of their origin languages in forming plural, singular and different verb forms.

The Persian language is written using modified Arabic script containing 28 Arabic letters to which 4 new characters were added (پ, چ, ژ, گ) - to express sounds not present in Classical Arabic. These 32 letters are written from right to left and for the most part of different forms according to their positions like whether they
are initial, medial or final, and connected or disconnected with the letter that precedes or follows them. This language does not contain the definite article in the strict sense (كتاب - means “book” or “the book” according to the context), however the particle which allows a definitive noun in accusative case (آب را بیاور - bring the water) can be said to perform the function of the article, so do the relative suffix (کتابی که ی - the book which) and ending (پسره - the son, informal writing). As for indefinite article it is expressed by suffix (کتا بهانی), placing a (hamze) over final - or single by means of numeral - “one”. There is no gender in Persian. The gender can be expressed by to use of different words (مرد - man, زن - woman), by means of the adjectives - “male” and - “female” and in the case of Arabic words the use of Arabic feminine ending (Dolamic & Savoy, 2009).

Plural of nouns in Persian language is formed by adding suffixes for animate beings (فداه - fathers) and for inanimate objects (گله - flowers). In some cases the plural ending can be written separately from the noun (e.g., خانه - or خانه ها - houses, first alternative being preferred). Arabic nouns take plural according to Arabic grammar by adding or for “sound” plurals or by alternating the vowel pattern of the singular for “broken” plurals (e.g. قلب – قلوب - hearts).

Suffixes predominate Persian morphology even though there are a few prefixes. Derivational Persian morphology is accomplished by means of prefix and suffixation of a stem, a usual construction with the Indo-European languages. Usually, the part-of-speech of the stem changes after adding a suffix (e.g., ’-ness’ in “good” and “goodness”) while the prefix changes the original meaning of the stem (e.g., “prehistory” vs. “historic” from the stem “history”). In Persian language this phenomenon occurs also with suffix adding. We can take as an example suffix - added
to nouns to from relative adjectives, resulting in ایران - ایرانی (Iran - Iranian), but also آب (Water-water).

### 3.2.1 Particularities of Persian Verbal System

One of the significant characteristics of the Persian verbal system is its small number of simple verbs. Actually most of the verbal concepts are expressed by compound verbs in this language. The syntactic and semantic features of Persian compound verbs have been the subject of interest for many linguists and some authors. Each compound verb in Persian is the combination of a nonverbal element and a light verb. The non-verbal elements which come before the light verb and in this sense are called the preverbal elements range over a number of lexical and phrasal categories such as noun, adjective, adverb and prepositional phrase.

The other interesting issue about the Persian compound verbs is the cause relation among them. Like English, Persian has lexicalized causative pairs but in contrast with English, the number of Persian causative pairs is very high. This fact results from a morph-semantic pattern among the Persian simple and compound verbs. Regarding the simple verbs, Persian has the suffix "–andæn" which can be replaced with the infinitive marker suffix "–idæn" and change into an intransitive, and causative verb to a simple transitive, causative one. The pair لرزیدان/ لرزاندæن (shake/shake) is of these kinds, you will find no cause relation between the two senses of the first verb (shake). Actually these two meanings are fused in one sunset and the definition "move or cause to move back and forth" shows that both causative and causative meanings are referred to the same lexical element and same sunset respectively. But regarding the corresponding Persian concepts, because we have two different lexical items we must construct two different sunsets and relate the causative one to the other by means of the cause relation.
One interesting point which causes a clear difference between English verbal sunsets and Persian one with respect to cause relation is that because in most of the cases in English there is no morphological realization for causation, this semantic relation is ignored and both causal and non-causal meaning are presented with one verb or sunset. For example \{close1\} is defined as "cease to operate or cause to cease operating" in Word Net 2.0 So in construction of their equivalent sunsets in Fars Net because there are two different lexical entries for both causative and non-causative meanings and we have made two different sunsets and linked one to the other by means of the cause relation.

It may be useful to divide the Classical Persian texts into two phases keeping the periods in mind, the early classical period (10th–12th century A.D.) and the late Classical period (13th–15th century A.D.). Due to the character of many of the Classical Persian texts, i.e. factual Prose such as historical chronicles, legends of saints, travelogues etc. with narrative passages mainly in the past tense there is only very limited expression of future to be found in them. Most occurrences of future, time reference are found in hypotactic constructions expressing conditions, temporal circumstances in the future, purpose, etc. and in modal constructions expressing necessity, volition, possibility, doubt, etc. Lexical or contextual means of expressing the future with the verb in the marked non-past form (Shamsfard, Hesabi, Fadaei, Mansoory, Famian, Bagherbeigi, & Assi, 2010).

3.2.2 History of Persian Language

As mentioned before, Persian language is one of the Indo-European languages. Indo-European languages are spoken by most of the people living in an area stretching from the Indian subcontinent to the eastern part of the Europe. They have also been
spoken in the United States since Europeans immigrated to this continent. These languages have originated from the original language, dubbed "mother" Indo-European, in very ancient times.

As its name suggests, the family of Indo-European languages is consisted of Indian and European languages (Safavi, 1988). It might be surprising for native speakers of Persian language to know that their language belongs to the same family to which English; French, German and Italian languages also belong, while Arabic language belongs to another family of languages. Let's take a look at the table below.

<table>
<thead>
<tr>
<th>English</th>
<th>German</th>
<th>Italian</th>
<th>French</th>
<th>Persian</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>father</td>
<td>vate</td>
<td>padre</td>
<td>pere</td>
<td>pedar</td>
<td>ab</td>
</tr>
<tr>
<td>brother</td>
<td>bruder</td>
<td>fratello</td>
<td>frere</td>
<td>baradar</td>
<td>ax</td>
</tr>
<tr>
<td>mother</td>
<td>mutter</td>
<td>madre</td>
<td>mere</td>
<td>madar</td>
<td>om</td>
</tr>
<tr>
<td>is</td>
<td>ist</td>
<td>e</td>
<td>est</td>
<td>ast</td>
<td>-----</td>
</tr>
</tbody>
</table>

Table 1: Comparative table of Persian and European languages.

From a comparative point of view, it could be easily found out that there are similarities between the five languages of English, German, Italian, French and Persian, while no similarities could be observed between these languages and Arabic language (Safavi, 1988).

Among the large family of Indo-European languages, the relationship and proximity between Persian and Indian is to such an extent that they should be examined by a single approach. Both Iranian and Indian people have introduced themselves as Aryan in their ancient works. For example, Darius the Great has described himself as "Aryan" and "from the Aryan race" in inscriptions dating back to his era. The terms "Eran" and "Iran", which mean the land of Aryan people, have derived from the word
"Aryan" later. In Indian language, the word "Aryan" is defined as people who spoke Sanskrit (Khanlari, 1968).

Iranian languages have been widely spoken in vast areas stretching to the Persian Gulf and the Sea of Oman in the south, Mesopotamia and Armenia in the west, and the Caucasus Mountains and the Deserts of Central Asia in the north. In the east, these languages were spoken all over Afghanistan and parts of the Pamir Plateau stretching to the mouth of the Indus River. Iranian immigrants once expanded their language to Northern Mongolia in the east and the western costs of the Black Sea in the west. Iranian languages are still spoken in a major part of this area (Yar-Shater, 1958). Iranian languages and dialects make up one of the most diverse categories of the Indo-European languages. A lot of research could still be done on these languages. Iranian languages and dialects could be divided into three categories based on their development processes viz; ancient Iranian languages, middle Iranian languages, and modern Iranian languages.

3.2.2.1 Ancient Iranian Languages

The ancient era is the oldest era of Iranian languages. Various Iranian languages have been spoken in this era out of which only four languages could be documented based on written documents available. Safavi,(1998) listed the four languages which the Median language, the Old Persian language and the Avestan language. (map, number, 1)

1) **Median language**: The median language has been spoken by Median kings as well as people living in western and central parts of Iran. "Median “people have been mentioned in the inscriptions of Assyrian kings since 835 B.C. Some Median words are also still found in Greek language, but most of our
information about this language has been obtained from words and expressions found in the inscriptions of Achaemenid kings-who followed Median kings (Moein, 1992). Among these words are ayiqayaxs (king), akarzav (great), nasa (stone), etc. which have been recognized as Median words according to linguistics rules.

2) **Sakan language**: Sakan was spoken by Iranian tribes living in vast area stretching from the costs of the Black Sea to the borders of China and northern part of areas settled by Parthians and Sogdians since the first millennium B.C. until the tenth century A.D. No written documents in Sakan language exists today, and we only know about four hundred words of this language, most of which are proper nouns. Some of these words are "atpah" (seven) and "adap" (foot).

3) **Old Persian language**: This language, which is related to Sanskrit and Avestan languages, has been the language of the people of Persia and the official language of Iran during the Achaemenid dynasty. The inscriptions of Achaemenid kings are the most important documents in Old Persian language that exist presently. The oldest inscriptions of this kind belong to Ariaramnes (640 B.C.), the great-grandfather of Darius the Great, while the latest inscriptions belong to Artaxerxes III (425 B.C.). The largest and most important inscription containing words of Old Persian language is the Behistun Inscription authored by Darius the Great on a mountain on Hamadan-Kermanshah route. About 500 words of Old Persian language could be found in this inscription, which has been written in cuneiform script. (Moein, 1992).

If we consider a sentence written in Old Persian language, for example, "OuhiehRadi" ("for that" or "for him"), and compare it with its Modern Persian
equivalent, we can find out that the Persian language has undergone more changes over time: five syllables have been reduced to two syllables, and all the words have been exposed to important changes not only in terms of the number of syllables but also in terms of their written shape. More over very little is known about the morphology and syntax of Old Persian language since sufficient Old Persian texts are not available.

4) **Avestan language**: Avestan language has been spoken by people living in eastern and north eastern parts of Iran. The holy book "Avesta" has also been written in Avestan language in different eras. The Avesta has been written in the Avestan script, which was extracted and developed from the Pahlavi writing system during the Sassanid era.

This language shares etymology with the Sanskrit language. According to linguistic studies, Avestan and Old Persian languages are two parallel and independent forms of a single language. In other words, Avestan language has been the dialect of a province while Old Persian language has been the dialect of another province. Determination of the date and duration of Avesta depends on research on the period of time when Zoroaster lived. No orientalist and researcher have so far managed to determine the exact date of Zoroaster's life. As a result, the exact date of writing Avesta remains unknown (Yar-Shater, 1958).
2.2.2 Middle Iranian Languages

The middle era starts from the fall of the Achaemenid dynasty and continues through the collapse of the Sassanid dynasty. This period of time has been determined through estimation since the oldest works belong to this era date back to about four centuries after the fall of the Achaemenid dynasty. Also, texts on this era date back to the middle of the tenth century A.D. (Safavi, 1988). Among the criteria for identifying the middle languages are the difference between these languages and their older versions (which are simpler generally), and their obsoleteness. The middle-era languages are divided into two groups based on their linguistic structure and geographical origin as eastern middle Iranian languages and western middle Iranian languages.

Western middle Iranian languages have been the official language of Iran during the Sassanid dynasty and regarded as the continuation of Old Persian language (Safavi, 1988). To Old Persian, Median, Middle Persian (Pahlavi), and Modern Persian languages belong to the western middle Iranian language group. Sogdian, Khwarezmian, and Ossetian (Ossetic) languages belong to the eastern middle Iranian language group. Avestan language is similar to the western group in some aspects and also similar with the eastern group in some other aspects. Therefore, it is not easy to assign it to one of these two groups. This language is one of the languages spoken in the eastern part of Iran.

The above categorization could also be applied to modern Iranian languages and dialects. The Persian, Kurdish, Luri, and Balochi languages as well as their dialects spoken in the southern coasts of the Caspian Sea, central Iran and eastern Iran belong to the western group, while Pashto and Yaghnobi (Neo-Sogdian) languages as well as
Iranian dialects spoken in Pamir and Ossetia (whose people immigrated to Caucasus) belong to the eastern group.

In the middle Iranian languages, the grammar has become simpler, the morphology of nouns, adjectives, and pronouns has been summarized, the detailed system of verbs with numerous participles, cases and tenses has been simplified, use of prepositions to determine different cases of the noun has become more common than ancient languages, and phrasal adjectives have become more popular.

In the western middle Iranian languages, the evolution has been towards simplification. In Pahlavi and Parthian languages, nouns are not conjugated, Deuteronomy is not observed, and determination of male, female and neuter is practically impossible. Eastern middle Iranian languages share traits with respect to words and grammatical rules which differentiate them from eastern middle Iranian languages (Yar-Shater, 1958).

Some of the middle Iranian languages are briefly introduced here:

1) **Parthian language (Parthian Pahlavi)**: this language, which has been spoken by Parthians who lived in north eastern Iran, was popular during the Parthian empire. Two types of works written in Parthian language currently exist: works in Parthian writing – derived from Aramic- and Manichaean works in Manichaean script –derived from Syriac alphabet. Among the dialects spoken in Iran, none of them could be described as a continuation of the Parthian language. Dialects spoken in Khorasan region today are Persian, and the original language spoken in this region has become extinct due to the invasion of different tribes as well as the influence of the official language of Sassanid era. But the Parthian language has influenced the Middle Persian language
during the Parthian empire and after that. Such an influence could be observed in Persian language today (Moein, 1992).

2) **Middle Persian language**: this language is the middle version of the Old Persian and the Modern Persian languages. It has been the official language in Iran during the Parthian empire. Out of the Middle Persian works remaining today, two works are mentioned here: 1- inscriptions of the Sassanid era written with an alphabet derived from Aramic but different from the Parthian alphabet and 2- Pahlavi books- most of which are Zoroastrian works written with an alphabet which is the continuation of the alphabet used in Pahlavi inscriptions (Yar-Shater, 1958).

It is noteworthy that Persian and Parthian are two dialects that are quite related to the Modern Persian language.

3) **Sogdian language**: this language has been spoken in Sogdiana, whose central cities included Samarkand and Bukhara. Once Sogdian language was the international language of the Central Asia and was expanded to China. Sogdian works have recently been unearthed in Central Asia and China. Sogdian alphabet has been derived from Aramaic alphabet in which Hozvareshes (words which were written in Aramaic and read in Pahlavi) are used, although the number of these Hozvareshes is low. Most Buddhist texts as well as non-religious texts and the Sogdian version of the Ghorbelgson Inscription, located in Mongolia, are in Sogdian alphabet. Christian works have been written with Syriac alphabet, and Manichaean texts have been in Manichaean script (Moein, 1992). Sogdian language gradually disappeared under the influence of Persian and Turkish languages. It has seemingly been spoken even through the twelfth century. Today, the only remnants of this language are the dialect of the people
of Yaghsub, who live in one of the valleys of the Zaravshan River (Yarshater, 1958).

### 3.2.2.3 Current Iranian Languages and Dialects

A variety of Iranian languages and dialects are spoken in areas stretching from Caucasus to Pamir and from Oman to Central Asia today. Among these languages and dialects, Persian is of special importance thanks to its rich literature and historic credit as well as its influence on other Iranian and a number of non-Iranian (Turkish and Urdu) languages and dialects (Yar-Shater, 1958). Current Iranian languages and dialects could be divided into several groups. They are primarily divided into two groups of eastern and western languages. Eastern Iranian languages that are still spoken are as follows:

1- Ossetia language, 2- Yaghnobi language, 3- Pashto language, 4- Pamir dialects.

The group of western Iranian languages is more diverse. The most important languages of this category are as follows: 1- Balochi language, 2- Kurdish language, 3- dialects of the Caspian Sea coasts, 4- Azerbaijani language and the dialects of Kor-Bangan in the south of the Aras River, 5- Southern languages and dialects.

The excerpts from Yar-Shater (1958) are briefed here.

**Ossetian language:** this language is spoken in parts of the mountainous areas of Central Caucasus. Ossetia language is of special status among current Iranian languages. This language is among a small number of languages that have not been influenced by Persian language. It has preserved many properties of ancient Iranian languages and is regarded as one of the oldest Iranian languages spoken today.
Ossetian literature includes interesting stories and fables recorded in the previous century. Among the most famous stories are a series of stories dubbed "Ghahremanan-e Nart" (the Nart heroes). Ossetian language is written with Russian alphabet today. Also, the language's derivations from Russian language expand day by day.

1- **Pashto language**: this language is the local language of Afghanistan and parts of the northwestern regions of India. The Pashto language spoken in Afghanistan today has drawn more attention currently, although it has been influenced by Persian and Arabic languages. It has preserved many original characteristics of Iranian languages and has different dialects. Pashto language enjoys literature and literary works as well. Among the poets of Pashto language is Khushal Khan. A large amount of local poems, stories and fables in Pashto are also available.

2- **Kurdish language**: Kurdish is a general term applied to languages and dialects spoken in the Kurdish areas of Turkey, Iran and Iraq. Some of these languages should be regarded as independent languages since they are too different from Kurdish (Kurmanji) to be versions of it.

The main Kurdish language is called "Kurmanji", which has numerous dialects such as Mokri, Sulaymniyah, Sanandaji, Kermanshahi, Bayazidi, Abdoyi, and Zandi.

3- **The Balochi language**: this language is spoken in Baluchistan and parts of Turkmenistan. Balochi is essentially categorized under the northern languages of the western Iranian language group. Baloch people have seemingly immigrated from the north to the south. Balochi language has borrowed from eastern Iranian languages because of proximity to them. This language has
different dialects, most important of which are Western Balochi and Eastern Balochi. Each dialect has its own sub-dialects, but there are not too many differences between these dialects and sub-dialects since Baloch tribes have always had relations with each other.

4- **New Persian Language (Dari):** this language is the most important Iranian language and dialect and is regarded as a continuation of Middle Persian and Old Persian, which originates from the language of Ancient Persia. New Persian represents the category of southwestern Iranian languages. This language has also been called Dari, Dari Persian, Persian and Farsi since the ninth and tenth centuries and after it became an official language following the establishment of eastern dynasties during the Islamic era.

New Persian was used by poets and authors after it became the language of the court and administration, with poets and authors starting to write and compose in New Persian gradually. After a while, great poets such as Rudaki, Daghighi, Ferdowsi and Kasayi produced great works in New Persian in the tenth century (Moein, 1992). It is clear that Dari dialect did not remain in its original form after it spread to all parts of Iran as an official dialect of politics and literature. The poets of the central, western, northern and southern parts of Iran lent some singular and phrasal terms of their own local dialects to Dari dialect. Also, Arabic language continued to influence this dialect. As a result, many singular and phrasal terms as well as conjugation and derivation rules of Arabic language penetrated into Dari dialect to form the Persian dialect that is spoken today. (Moein, 1992).

The development of the current Persian language began with the first works written in Persian in the post-Islam era. Important social, political and religious developments that took place in Iran after Islam entered the country affected Iran's
official language, which was recorded with Arabic alphabet rather than Pahlavi alphabet in the post-Islam era. The oldest Persian texts written with Arabic alphabet are the poems of Hanzala Badghisi, Mohammad Ebn-e Vasif Sistani, Shahid, and Rudaki. The current Persian language is a continuation of Middle Persian in terms of grammar (Yar-Shater, 1958).

New Persian has turned from a language with special restrictions in morphology and syntax into a very simple and analytical language, getting rid of severe conjugation restrictions observed in ancient Iran. However, this language has managed to meet the same goals it fulfilled in previous eras through the use of new forms of verbs and propositions. Today's Persian language is a continuation of Middle Persian in terms of grammar (Moein, 1992).

3.2.3 Persian at the Mughal Court

Persian at the Mughal Court to what extent Ahmad Kashii and Safi ud-Din, or for that matter the other scholars to whom Akbar chose to write in Persian, could extend his influence in Iran is a matter of debate. They were executed by the Shah together with many others like them. However, the Mughal emperor's desire of bringing 'the exalted [Iranian] community closes to him spiritually and materially' prepared the grounds for many of them to make India their second home. The Iranian talents flourished more in Mughal India. Surely, Iranians never accepted Akbar's claim of being 'the solver of their spiritual problems' which pertained to the life hereafter; soon, however, the belief in literate Iran was widespread that a visit to India promised material comforts and honoured position in this world.

In the process India drew close to Iran culturally, and this, among other things, helped Persian attain the status of being the first language of the king and the court in
Mughal India. Among the first literary works of the reign of Akbar, at a time when he was consolidating Mughal power in India was the preparation of a Persian translation of Babur-nama. Ironically, the translator was 'Abdur Rahim Khan, Khan-i Khanan, the son of Bairam Khan, who was also a poet in Turkish. But it was not simply that Babur’s memoir was to be rendered into Persian, the Emperor also desired that the sources of the new court history recording Mughal achievements be compiled in Persian. Humaytin's sister, Gulbadan Begum's Humayan-nama, was written in Persian, even though Turkish was the native tongue of the Princess and her husband, Khizr Khwaja Khan. Indeed, Mrs Beveridge, who translated Gulbadan's account into English, suspects that the book was originally composed in Turkish. Akbar did not have any formal education. Important books were therefore read out to him regularly in his assembly hall. His library consisted of hundreds of prose books and poetical works in Arabic, Persian, Hindi, Greek and Kashmiri. But the books which the Emperor heard repeatedly were all in Persian (Alam, 1998).

Persian thus emerged as the language of the king, the royal house-hold and the high Mughal elite. Akbar's son and successor, Jahangir (1605-26), was not good in Turkish, but he had his own style in Persian and wrote his memoir in an elegant prose. He was also a good critic of Persian poetry and composed several verses and ghazals. It was for him that Jayasi’s Padmdvat was translated into Persian, but the work was recognized only as an Indian fable (afsana-i Hindf) and not as a book on Islamic mysticism in Hindi. Still later, with volumes of letters and edicts, Aurangzeb (1656-1707) established himself as a fine prose writer of his time. The formal abolish of the institution of malik-ush-shu'ard' affected a little of the supreme status of Persian. Indeed, later seventeenth-century northern India witnessed numerous native poets of
high standard in Persian, including the great Mirza 'Abd-ul-Qadir Bidil (d. 1719) and Nasir Ali Sirhindi (d. 1696) (Alam, 1998).

3.2.4 Some Points about Persian

Before dealing with the actual tagset, some basic facts about Persian seem to be in order. Persian is an Indo-European language belonging to the Indo-Iranian branch. The grammar of the language has been largely simplified since Old Persian and the vocabulary has been greatly influenced by Arabic and to a less extent by French. Only verbs in this language are inflected. There are fewer tenses in Persian than in English; however, Persian enjoys a widely used subjunctive mood. Persian is an SOV language and direct objects are marked by its only postposition rā, which also functions as a topic marker in spoken language. There are disputes over the functions of rā. Therefore, we shall resort to the traditional term “direct object marker.” However, to show that we do recognize that it has other functions as well we have placed it in the category ra whose only member is the postposition in question.

The language does not make use of gender even him /her distinction that exists in English. Only nouns are made plural in this language and even uncountable nouns can be made plural. Possessiveness is indicated through the genitive morpheme –e, which is invisible in writing. The existence of a direct object marker enables speakers of Persian to use subjects and objects in a free order although the standard usage is SOV. Adverbs appear virtually everywhere. Adjectives mostly follow the noun they modify but there are some compounds in which adjective preceded the nouns. Verb and noun phrases are the most consistent phrases in terms of word order and this enabled us to use more tags that correspond to members of these phrases. (Abdolhosseini, 2000).
3.2.5 Iranian Persian vs. Indian Persian

Amir Khusrau (1253-1325), with whose poetry and writings the Indian diction assumed a distinct personality of its own, had disapproved of the 'Khurasanzf' idiom and had noted that in India Persian was written and pronounced according to the standard of Taran. The models in particular in prose then were the authors of Transoxiana. The writings of Rashid-ud-Din Watwat and Baha-ud-Din of Khwarizm, for instance, were read and imitated by the Indian Persian writers. Khusrau then innovated a new Indo-Persian style. To the pre-Mughal lexicographers the languages spoken in Shiraz, Mawara-an-Nahr and Farghana were just dialects of the same Persian tongue. They incorporated in their lexicons the words used in 'Fars, Samarqand, Mawara-an-Nahr and Turkistan'. They also give as a matter of routine their Hindavi synonyms. The Mughal court, on the contrary, was concerned with the purification of Persian (tathir-i Farsf). The objective of the lexicon Akbar asked Jamalud-Din Husain Inja to prepare in his name was to clean the Persian of non-Persian words and expressions. The drive at purification continued later. Inja's Farhang together with Majmaul Furs Surari were taken as the sole standard lexicons in the first half of the seventeenth century. By the middle of the century Mulla 'Abd-ur-Rashid Thattawi felt the necessity of compiling a new dictionary, among other things, because(a) in the existing two major dictionaries (Jahangiri and Surari) certain Arabic and Turkish words were enlisted without clarifying that they were not Persian, and that (b) many words were wrongly pronounced by Inja and Suriri.

If in the late sixteenth and seventeenth centuries the most sublime and the best poetry of the time were produced in India, the eighteenth century was the richest in terms of the number and varieties of prose and poetic works in Persian. Take, for instance, the poetry, seventy-seven of the Persian poets who lived during the earlier half
of the century found the place of honour in the tazkira and titled as Majma'-un-Nafa'is of Siraj-ud-Din. Among them 'Ali Khan Arza was the best and the most revered scholar, poet and critic of the period.

Many tazkiras were written in this period. Among these, 'All Ibrahim Khalil's (d. 1793) Suhuf-i IbrahItm, which was the most comprehensive one, at least, for the northern Indian poets, notes no less than four hundred and sixty poets from amongst those who lived in the eighteenth century and whose works he considered of any worth. No less than fifty-six of them were non-Muslims. Much more significant was the height of excellence the Persian poetry scaled under the Mughals. In Central Asia, at the court of the late Timurids, Daulat Shah Samarqandi and 'Ali Sher Nawa'I tried to establish a canon which, in a measure, was represented in the poetry of 'Abd-ur-Rahman Jami (d. 1492) and which later was refined and reformulated by Babur. In this tradition emphasis was on rhetorical artistry, even as Babur pointed to the importance of meaning (ma'nt) and ecstasy (hal), together with colour (rang), in a good poetry. In Iran, too, there were attempts like the one by Sam Mirza to evolve the standards of literary criticism, close to Herat School. Simultaneously, however, Baba Fughani Shirazi (d. 1519) made a plea Poetry and the poetic imagination to Faizi was something beyond this world. The poet was to scale the heights, insurmountable for an average human soul. For concerns in the poetry for the routine matters of love, but with an eye for new meanings of old words.

Throughout the Mughal age, the poet thus showed awareness of the fact that the realm of the new poetry expanded much beyond its erstwhile frontiers. In relation to their predecessors, they saw themselves collectively as original, creative and innovative, even though each of them diverged from the other and experimented with
new images and tropes in his/her own individual style. Enthralled with the poetry of his own time, the Mughal poet is obsessed with a sort of collective ego. Mirza 'Abd-ul-Qadir Bidil (d. 1712), for instance, had not much in common with 'Urfi, and yet he seems to be conscious of the fact they both belong to the same group.

True, there were some attempts for Indian Persian to acquire an autonomous position, but these were feeble and exceptional. In fact, in a bid to legitimize the use of the Indian words in Persian he earned the distinction of being the first to discover and point out the correspondence between Persian and Sanskrit. However, the choice of Persian, as the language of the empire was, in a very large measure, also in consideration of the specific Indian conditions. The non-sectarian and liberal feature of Persian made it an ideal forum through which the Mughals could effectively negotiate the diversities of the Indian society. The culture and the ethos of the language matched with their vision of an over-arching empire. Persian became a particularly useful instrument for political maneuverability, also because pre-Mughal India had developed familiarity with the language (Alam, 1998).

3.2.6 The importance of Persian Language

Persian language is one of the greatest Iranian languages that are used in speaking and writing today. Among numerous Iranian languages, Persian has reached a status in Asia that no other Iranian language has ever managed to reach. In the past, Persian was spoken in a vast area stretching to Middle Asia in the northeast, China in the east, the whole Indian subcontinent in the south, and the whole Asia Minor in the west. It was the literary and administrative language in India for centuries.

Persian is the official language in Iran and Afghanistan (where Pashto is a second spoken language) presently. It is also spoken in neighboring countries such as
Pakistan, India, Iraq, Tajikistan, Caucasus (Azerbaijan). Persian literature, especially its romantic and mystic poems, is among the best in the world. Persian also has a significant status in Indology studies. It has shown its flexibility and development in general linguistic studies. What has remained of the old form of Persian could be analyzed. As Moein, 1992 rightly mentioned that Persian is also of great importance in terms of influencing Arabic, Turkish and Urdu languages as well as European languages (French, English, Russian, etc)

### 3.3 Persian Linguistics

Each human language is a complex of knowledge and abilities enabling speakers of the language to communicate with each other, to express ideas, hypotheses, motions, desires and all the other things that need expression. Linguistics is the study of these knowledge systems in all their aspects and analyze further how is such a knowledge system structured, how is it acquired, how is it used in the production and comprehension of messages, how does it change over time?

In the last thirty years, linguistic research has gone beyond the study of the internal structure of a language and overlapped with other fields like neurology, psychology, computer science, sociology, and formed new interdisciplinary fields such as neuro-linguistics, psycho-linguistics, computational linguistics, and socio linguistics. Although it is less than a century since linguistics was deemed a new science, its development has been very fast and remarkable. Different research and investigations on Indo-European and non-Indo-European languages played a major role in the development of this new field. Persian is one of the Indo-Iranian languages that attracted the attention of many linguists who studied its various linguistic aspects. The Persian linguistic research began with the study of the grammar and the subsequent
writing of grammar books. Later, these studies expanded to the lexicography, writing and editing of bilingual dictionaries.

3.3.1 History of the Study of Grammar in Iran

Modern Persian, which developed from Middle Persian (ninth century), is an Eastern Iranian language and is spoken in Iran as Farsi, Afghanistan as Dari, and Tajikistan as Tajiki. The study of grammar in Iran is attested as early as the Sassanian dynasty (224–651 AD). Even then phonetics (the study of sounds), morphology (study of word formation) and the lexicon (the abstract information that a native speaker of a language has in mind) were being studied in order to preserve the sacred Zoroastrian texts. In the Islamic era, Iranians pioneered writing grammars of Arabic.

3.3.2 Linguistics Works from Seventh Century to Ninth Century

Following the Arab invasion, the method used by grammarians in the earlier centuries was prescriptive. The main interest was based on the form of words and sentences rather than a general pattern which held for similar cases. When Muslims learned Greek logic and became familiar with grammars written for Greek, their methods changed. The grammarian started to categorize and formulate general patterns in a logical order. The most important grammars written by Iranian scholars during the first centuries after the Arab invasion were on Arabic.

As Khalil Saman indicates in his linguistics in the middle ages, “the earliest complete analysis of the Arabic language still available to us is that of Sibaʾwai, the classical master of the Arabic linguistic tradition.” Alkitaʾb (the book) is the title of Sibaʾwai’s famous work in which he incorporates the phonetics of Arabic. He also discusses the place and manner of articulation of sounds. The second Persian grammarian was Zamakhshari and his work on Arabic grammatical principles is
Almufassal. The phonetic part of Almufassal is very similar to the Alkitab’s phonetics. However, Zamakhshari discusses the principles of Arabic grammar in more detail than his predecessor. The first scholar to discuss Persian grammar was Avicinna (Ebn-e Sina, 980–1037 AD) in his Da’nesha-me-ye Ala’i. Avicinna attempted to show the universality of language and how it functions to express the logical hierarchy of the “essential.” Avecinna’s approach to grammar was beyond the pragmatic descriptive/prescriptive view with which Arabic grammar was investigated and was similar to structural linguistics of the sound system, word formation principles and grammar of Persian languages from Old Persian to Modern Persian.

Among non-Iranian scholars of Old Persian and Middle Persian, Robert Stephen Beekes whose 1988 Grammar of Gatha-Avestan is renowned among Old Persian grammars. Roland Kent also presented a detailed description of Old Persian lexicon, texts and grammar. Other non-Iranian scholars of Middle Persian languages are Christopher Brunner who worked on the syntax of western Middle Persian Languages and Ilya Gershevitch who provided a grammar on Sogdian. The first systematic study of the stages in the development of Modern Persian was done by Bahar. His three volumes of Sabk Shenasi cover and combine the grammatical and rhetorical/poetical aspects of Modern Persian. In addition to the well-known scholars who effectively contributed to the development of diachronic aspects of Persian linguistics, there are some governmental institutions which have an important role in diachronic and we will see later in synchronic Persian linguistics. One of these institutions is Farhangesta’n-e Ira’n (Academy of Iran). Farhangesta’n-e Ira’n made many efforts to revive the intrinsic power of Persian to create new words for concepts. These efforts have been continued by the Farhangesta’n-e zaban va adab-e Fa’rsi (The Academy of Persian Language and Literature). The main goal was to create words
either by revitalizing obsolete words or by revitalizing the derivational processes of Persian. Among the scholars who worked in this field are Kasravi, Moqadam, Khanlari, Bateni and Milanian (Taleghani, 2009).

3.3.2.1 Diachronic Linguistic

Diachronic literally means *across-time*, and it describes any work which maps the shifts and fractures and mutations of languages over the centuries. In gross outline, it is similar to evolutionary biology, which maps the shifts and transformations of rocks.

It was only in the 19th century, however, that more scientific methods of language comparison and sufficient data on the early Indo-European languages combined to establish the principles now used by historical linguists. The theories of the new grammarians, a group of German historical linguists and classical scholars who first gained prominence in the 1870s, were especially important because of the rigorous manner in which they formulated sound correspondences in the Indo-European languages. In the 20th century, historical linguists have successfully extended the application of the theories and methods of the 19th century to the classification and historical study of non-Indo-European languages. Historical linguistics, when contrasted with synchronic linguistics, the study of a language at a particular point in time, is often called diachronic linguistics.

3.3.2.2 Synchronic Linguistics

Synchronic linguistics is one of the two main temporal dimensions of language study introduced by Swiss linguist Ferdinand de Saussure. It concerns with the events or phenomena at a particular period without considering historical antecedents.
Persian synchronic linguistics studies can be classified into three main stages. As formulated by Taleghani, (2009), 1) traditional prescriptive linguistics, 2) structural linguistics, and 3) generative or formal linguistics. We will see how each one of these stages affects Persian linguistics research and studies.

### 3.3.2.3 Traditional Prescriptive Linguistics

The grammar which lays down rules determining whether or not a particular arrangement or use of language is ‘correct’. It is ‘prescriptive. We say that traditional grammar is prescriptive because it focuses on the distinction between what some people do with language and what they ought to do with it, according to a pre-established standard. The chief goal of traditional grammar, therefore, is perpetuating a historical model of what supposedly constitutes proper language.

In traditional grammar, the basic word classes were classified as nominal, verbal, and particle, each being further subdivided into smaller classes. The main logical elements were subject and predicate. Persian grammars distinguished further actional sentences with a verb, from stative sentences consisting of an initial statement (subject) and the affirmative/predication by a substantive (verb) phrase.

### 3.3.2.4 Structural Linguistics

Structural linguistics began in Iran with Bateni’s book Towsif-e Sa’xtema’n-e Dasturi-ye Zaba’n-e Fa’rsi (Structural Description of Persian Grammar). This book, published in 1967, opened a window for Iranian linguistic circles to examine the new approach to linguistics. Bateni presents the basic paradigms of the Persian sentence, clause, phrase, and word in his book on structural grammar. Another Persian grammar on the structural linguistics approach was Martine’s Ele’ments de linguistique Ge´ne´rale´, which was used as a textbook by Milanian for introductory courses on
linguistics for a decade. Other non-Iranian scholars who presented Persian grammar in the frame of structural linguistics were Lambton in 1953, and Lazard in 1992. We can see the effect of structural linguistics in the Persian phonetic and sound system.

The first book on general phonetics was written by Haghshenas in 1977. He in his book Towsif-e a`va`yi-e zaba`n-e Fa`rsi (Phonetic Description of Persian), which was published in 1978 given a more elaborated on were Haghshenas with a description of phonotactic principles and syllabification of the Persian sound system (Taleghani, 2009).

3.3.3 Generative and Formal Linguistics

Generative linguistics is a school of thought within linguistics that makes use of the concept of a generative grammar. In the mid 1970s, Persian linguistics turned toward formal or generative linguistics of American. Bateni was the first Iranian linguist who familiarized among Iranian linguists, and also published his New Approach to the American Grammar of Language in 1977. In this book, he introduces the principles of American structural grammar, as well as the theory of transformational generative grammar. Formal or generative linguistics became more popular as many Iranians began to train in American universities. The first typological grammar of Persian written in the frame of generative linguistics is Dasture zaba`ne Fa`rsi az didga`he Radeshena´si (Persian Grammar: A Typological Approach) by Shahrzad Mahutian (1999). This grammar is a synchronic descriptive grammar of the Tehrani a spoken dialect of educated people. It contains three parts: 1) morphology and word formation, 2) syntax/grammar, and 3) sound and sound system. This typological grammar compares some of the features of Persian grammar with other languages.

Among the first dissertations written on Persian grammar and syntax in the frame of transformational grammar were those by Moyne (1974) Mara`shi (1970) and
Bashiri (1972). The most important role of these dissertations was to create a revolution in Persian grammar studies and also to herald the arrival of theoretical linguistics in the field of Persian linguistics in the twentieth century. They attempted to analyze different aspects of Persian grammar within the framework of formal linguistics. Among them the work of Soheili, whose 1976 dissertation focused on the noun-phrase complementation as formal linguistics theory developed by Chomsky and others in the 1980 and 1990s. The Iranian linguists in and out of Iran changed the perspective of Persian linguistics and followed the new trends in the field. Therefore, we saw a lot of developments in syntactic aspects of Persian linguistics during those two decades.

Linguists who made a great progress in syntactic aspects of Persian include Samiian, Hashemipour, Dabir-Moqadam and Karimi. Samiian was the first Iranian linguist to give an analysis of noun phrase and Eza’fe construction in Persian. Hashemipour’s work focused on the specific category of Persian verbs and Dabir-Moghadam concentrated on causative constructions in Persian. Among all the Iranian linguists who worked on Persian in the 1980 and 1990s, Karimi is the only one whose research includes a wide range of morpho-syntactic aspects of Persian language. She started with specificity and definiteness in Persian in 1989 then gave a detailed analysis of the object and definite marker ra` in 1996. She followed this with Persian complex predicates in 1995 and 2005, which constructions in Persian in 2001 and 2005 and many other articles in different areas of Persian syntax. Her book on Persian syntax was the first one written in this field and her contribution to Persian linguistics is not limited to her research.

She was the first Iranian linguist to establish and organize two workshops and conferences on aspects of Iranian linguistics with the cooperation of Vida Samiian and Don Stilo. A volume on aspects of Iranian linguistics published in 2008 including some
of the papers that were presented at the conference in Leipzig in 2005. These conferences are a good opportunity for scholars who work on different linguistic aspects of Iranian languages to share their research with others.

As the research on Persian linguistics and the study of Iranian languages developed so remarkably and quickly a need was felt for a journal dedicated to presenting and sharing theories and ideas. Karimi proposed to start working towards establishing an Iranian linguistics journal in summer 2008 at the second workshop on aspects of Iranian linguistics in Germany.

### 3.3.4 Current State of Persian Linguistics

Following the developments of Persian linguistics in the 1980s and 1990s, Iranian linguists continued their work on different aspects of Persian linguistics in Iran and out of Iran, from Iranian universities to European, American and North American universities. The field of Persian linguistics in North America is represented by Jila Ghomeshi at the University of Manitoba. Ghomeshi graduated from the Linguistics Department of the University of Toronto in 1997. Her major interest in linguistic research is the syntax of Persian. She edited a volume on Iranian linguistics for the Canadian Journal of Linguistics (The Syntax of Iranian Languages), and she has contributed articles on various aspects of Persian linguistics including “Plural Marking, Indefiniteness, and the Noun Phrase,” and “Non-projecting Nouns and the Eza’fe Construction in Persian.”

Among the new generation of Iranian linguists working on formal linguistics the notable scholars Kahnamuyipour, Karine Megerdoomian and Azita Taleghani. Arsalan Kahnemuyipour graduated from the Department of Linguistics at the University of Toronto in 2004. His research focuses on the syntax of sentential stress. Karine
Megerdoomian graduated from the Department of the Linguistics at the University of Southern California in 2003. Her main research focuses on Persian complex predicates. Azita Taleghani concentrates on modality, aspect, tense and negation in Persian, graduated from the Department of Linguistics at the University of Arizona in 2006, and it is worth noting that there are many young Iranians in the United States, Canada, Europe and Iran who are currently working in the field.

The valuable contribution of linguistics is evident in various educational organizations in Iran. Among them we can name the Academy of Persian Language and Literature, the Iranian Cultural Heritage Organization, the Iranology Foundation, and the Educational Ministry. In the twenty-eight years since the revolution, the Academy of Persian Language and Literature has become one of the most prominent institutions in Iran taking steps in linguistic studies from diachronic studies of Iranian languages to etymological investigations, and lexicography.

The main activity of this organization has been creating and approving official Persian equivalents for general or technical terms of foreign languages. It has also created an official orthography of the Persian language, titled Dastur-e khatt-e Fa`rsi (Persian Script Orthography). The Academy also promotes research works on various aspects of Persian language and literature, Persian heritage, Iranology, other Iranian languages, and Iranian culture.

Following the 1970s trend in revitalizing and creating new words for obsolete ones, linguists in the komite-ye va`jegozini (word-formation committee) of the Academy of Persian Language and Literature have worked with the scholars of different sciences to create many new terms based on the principles of Persian word formation for foreign scientific terms in different sciences. This research unit of the
Academy has been very productive in the last 15 years, creating many terms for more than 30 scientific fields. Linguists in the Academy of Persian Language and Literature are not only involved in creating new equivalents and on official orthography of Persian but also engaged in the process of writing a descriptive grammar for Persian.

The Iranian Cultural Heritage Organization is another institute in Iran which following the revolution played a major role in the study of diachronic aspects of Iranian languages, as well as ethnographical and anthropological research. In the last ten years, the Educational Ministry of Iran with the cooperation of Iranian linguists has taken several steps toward changing the standard grammar and textbooks of literature. Recently, the linguists in the Educational Ministry started a new project on the Basic Words of Persian. It is expected that the results of this important project will be useful for studies in applied linguistics, first and second language acquisition and educational investigations further.

The new Iranian Linguistics Association which was established in 2004 through the efforts of some of the Iranian linguists and students of linguistics in Iran. An Institute for Cognitive Science Studies (ICSS) was established by scholars of psychology and cognitive sciences in Iran in 1998.

The governmental organizations are not the only institutions that affect Persian linguistics. Some private publishers, such as Farhang-e Moa’ser, in cooperation with linguists and expert lexicographers, have published dictionaries and encyclopedias and played a prominent role in the lexicographical research in Iran. Therefore the development of Persian linguistics in the twentieth century can attributed to all these associations established in the last ten years.
3.3.5 Future Prospects for Persian Linguistics

Despite all these efforts, the Persian linguistics has a long way to go as visualized by the experts. They also felt that the Persian grammar still needs a good descriptive standard text. Therefore, it is the part of the linguist’s duty to cooperate with and encourage the grammar section of the Academy of Language and Literature of Iran to finish this project. Similarly the Persian language instructors have to develop various textbooks for teaching Persian at different levels. This can be possible with the help of Persian language instructors, applied and theoretical linguists and one has to keep the language situation also mind as Persian is a language with many dialects. The structural investigation of these dialects will help the academics in linguistically and theoretically to achieve the objectives.

3.4 Corpus in Persian Lexicography

"A corpus is a collection of texts or part of texts upon which some general linguistic analysis can be collected" (Meyer, 2002). It is applied to a collection of written or spoken language data that could be employed to describe and analyze a language. Undoubtedly, any research on languages and their description and analysis, including lexicography, grammar compilation, dialectology and alike, will not yield useful results without relying on corpora. The term "corpus" is especially used by structural linguists, who always emphasize that the description of a language or dialect should be based on collecting and analyzing data.

Many linguists have emphasized the importance of corpus in most of their researches in the past. However, relying on real linguistic data has become common in the modern era and has turned into one of the essential requirements of many theoretical and applied studies such as studies on dialectology, grammar compilation, lexicography and language teaching and testing. Thus, a specialized branch in the field
of computational linguistics has emerged. This new branch created into 20th century and named as corpus linguistics became one of the most active branches of linguistics within a short period of time.

3.4.1 Resources (Components)

Corpus is the most important factor contributing to compiling a comprehensive monolingual general dictionary. According to Hashemi Minabad (2007), the components of a corpus are as follows:

1- Written document, (from the oldest texts to the era of compiling dictionaries, including books and pamphlets on literature, religion, and history, legal and judicial documents, agreements, letters, stories, itineraries, magazines, newspapers, classified advertisements, etc.)

2- Spoken document, (telephone conversations, daily conversations, radio and TV debates, face-to-face interviews, classroom discussions, personal quotations, etc). Lexicographers also extract words, linguistic elements, and information related to spoken language from native speakers.

3- Document obtained through field research: (these documents are especially useful for compiling terms used in different professions and jobs.)

4- Linguistic knowledge of the lexicographer

The lexicographers are also a source of information in to compilation of dictionaries. The linguistic knowledge of lexicographers and their lexicographical skill guarantees the success of dictionary projects. The lexicographers with critical and analytical insight are capable of removing shortages in extracted data and correcting flaws observed in such data. They can compile good dictionaries with their knowledge about written language and spoken languages as well as stylistic, contextual, semantic and grammatical nuances.
Corpus can be drawn from the following sources:

**3.4.1.1 Written Corpus**

Written corpus is regarded as the most important linguistic corpus. Even dictionaries that are claimed to emphasize spoken language greatly benefited from written corpora. For example, Ahmad Shamlu who has allocated his "Street Dictionary" on vulgar language has quoted written documents in many cases. One of the main reasons behind the great emphasis on written documents is the availability of such documents compared to spoken documents. The holiness and importance of written language is another reason. Written corpora are of different types which are used depending on the type and audience of the dictionary. Written documents could include textbooks, fiction works, non-fiction works, newspapers, publications, previous dictionaries, and reference books.

**3.4.1.2 Textbooks**

Textbooks of various education levels are among the main sources of the corpus of a general dictionary. These textbooks enjoy great vocabulary diversity, cover many basic topics, and almost include all important fields. However, some terms found in such books are too specialized to be used in general language. For example, some very specialized terms found in the "carpet-weaving" in a textbook of the first grade of high school include "Shakheh-Pich Toranjar" (twisted branch with citrons), "Talfighi Toranji Dasteh Gol" (combination of citrons and flowers), "Talfighi Toranji Kaf Sadeh" (combination of citrons and simple backgrounds), "Talfighi Lachak Toranj Kaf Sadeh" (combination of citrons, scarves and simple backgrounds).

It is obvious that the textbooks contain many words used in different professions and fields. So, the language used to describe and define words in dictionaries should be
closely related to the language of textbooks. In addition, textbooks could also contribute to the language and expression of dictionaries. The corpus of general dictionaries does not cover university textbooks since they are too specialized. However, general textbooks such as "General Persian", "the History of Islam", "Islamic Insight", and "Physical Education" could be covered by the corpus of general dictionaries.

3.4.1.3 Fiction Works

Novels, stories, short stories and plays make up a major part of written corpus since fiction works reflect a wide range of human experience and are widely read. Literary texts and classic works often make up the main source of Persian dictionaries. Fiction works have different style and types, covering various topics. Historical, science-fiction, adventure, police records, entertainment and general stories can be proportionally included in written corpora.

3.3.1.4 Non-Fiction Works

More attention is paid to non-fiction works today. Non-fiction works include books on general, scientific and technical topics, itineraries, biographies, guide books, and collections of articles. These works cover different fields such as arts, culture, civilization, economics, history, politics, psychology, sports, nutrition, health, housekeeping, cooking, animal life, environmental health, industry, education, religion, cinema, etc.

3.4.1.5 Newspapers and Publications

Newspapers and periodical publications contain many modern and contemporary linguistic elements, reflecting the latest linguistic and non-linguistic developments. Newly-coined words, definitions of new words and existing terms, and
colloquial elements are widely found in such texts. An advantage of newspapers is that they cover various topics and their language is a current daily language.

3.4.1.6 Letters and Catalogues

Formal, administrative and personal letters have their own specific words that should be recorded by general dictionaries. Catalogues, brochures and classified advertisements often contain numerous and various elements of the contemporary and daily language which are rarely found in other texts.

3.4.1.7 Previous Dictionaries

The general dictionaries published earlier are one of the important sources of data. Sometimes, a dictionary is compiled based on another one. For example a concise dictionary is compiled based on a comprehensive one or vice versa. Quoting previous dictionaries to write a new one is a used practice and has long history. Therefore the data provided by other dictionaries could be employed to review the data collected for the new dictionary. However, this should be done with lot of case and researcher.

3.4.1.8 Reference Books

Reference books, encyclopedias, and dictionaries on language difficulties, such as *let's write correctly* by Abolhasan Najafi, are among the resources that can be used to describe and define words and extract data.

In the past for preparing linguistic corpora to compile a dictionary was a time-consuming, laborious and costly task. For example, Allameh Dehkhoda studied the Persian texts in verse and prose for many years and then prepared a trove of notes containing words, evidence, and references. The collection of linguistic data which Dehkhoda prepared through the traditional note-taking method has been unique.
3.4.2 Corpus in Persian Dictionaries in the Indian Subcontinent (13th – 19th century)

The lexicography became significantly popular in the Indian subcontinent in the twelfth century, after Islamic dynasties such as Ghurid and Mogul came to power in the subcontinent and Persian language was spoken in the Indian court. The Ghavas Dictionary by Fakhroddin Mobarakshah Ghaznavi was the first dictionary compiled in that era. The Loghat-e Fors dictionary, rather than texts, made up the main corpus of this dictionary.

It became necessary to compile more complete Persian dictionaries in the Indian subcontinent after great Persian-speaking scholars such as Amir-Khosro Dehlavi, Sa'ebTabrizi, and Bide Dehlavi emerged there. The Jahangir Dictionary was written by Mir Jamaloddin Hossein Ebn-e Fakhroddin Hasan Anju Shirazi during Mogul King Jahangir. It was regarded as an important evolution in Persian lexicography. The corpus of this dictionary included:

1- Previous dictionaries, including 53 dictionaries compiled to that date.
2- Texts in verse. Only one text in prose has reportedly been among the corpus of this dictionary (the NaserKhosrow's Itinerary).
3- Pahlavi words. The author has recorded and defined a number of Pahlavi words, known as "Zand" and "Pazand" words among lexicographers, in his dictionary with the help of the minority Zoroastrian population living in India, known as "Persians".
4- Other references such as Zakhireye Khwarazmshahi, Nozhatol Gholub, and Ajayebol Boldan (Khatibi, 2007).

A new chapter began in Persian lexicography and the corpus of Persian dictionaries and words increased following the compilation of the Borhan-e Ghate
Dictionary by Mohammad Ebn-e Hossein Khalaf Tabrizi. The corpus of this dictionary included:

1- Previous dictionaries and poetical works
2- Arabic dictionaries
3- Pahlavi words known as "Zand" and "Pazand"
4- Dasatiri words, as the words coined by Azar Kayan in a book named Dasatir-e Asemani during the reign of the Mughol emperor Akbar and attributed to a false prophet named Sasan the Fifth (Khatibi, 2007).

The compilation of dictionaries continued even after the Mughol Empire and during the British rule in India. Among the most important dictionaries was the Anenderaj, which used a relatively large number of words from Arabic dictionaries as its corpus. The last important Persian dictionary authored in India was the Nezam Dictionary by Mohammad Ali Daeeoleslam which has been compiled based on a new method. The corpus of this dictionary was essentially different from those of previous dictionaries:

1- The corpus of the dictionary included not only complicated poetical words, but also spoken words.
2- It has paid attention to text in both verse and prose.
3- "Zand" and "Pazand" words have been omitted.
4- Words and their spellings have been collected through referring to people.
5- Tehrani dialect has been employed to record words (Khatibi, 2007).

3.4.3 Corpus in Contemporary Lexicography in Iran

The dictionaries were also compiled in Iran from time to time concurrent with the compilation of Persian dictionaries in India from all of the sixteenth century dictionaries authored in Iran were influenced by those written in India. The Majma el-
**Fors-e Soruri** was one of the oldest and most important dictionaries authored in Iran during that period. The dictionary, compiled during the reign of Shah Abbas I, was totally based on the *Jahangiri Dictionary*.

In the contemporary era and concurrent with the evolution of Persian language, the basic changes started to occur in the corpus of Persian dictionaries. For example, the corpus of the *Nafisi Dictionary* by Nazemol-AtebbaNafisi included the European words (e.g. words from Persian-English dictionaries by Richardson, Wollaston, Johnson, and Steingass) in addition to words from previous dictionaries both in Persian and Arabic.

3.4.3.1 Corpus in the Dehkhoda Dictionary

*The Dehkhoda Dictionary* is regarded as the last Persian dictionary written based on traditional methods. It is also a collection of all the dictionaries compiled previously and covers almost all the words found in important handwritten and printed Persian and Arabic dictionaries. In addition, it contains thousands of sarcastic words and phrases as well as examples from prose and verse texts and other scientific and literary books none of which could be found in other Persian and Arabic dictionaries. *The Dehkhoda Dictionary* also covers many Turkish, Mogul, Indian, French, English, German and Russian words which were used in Persian but had not been included in dictionaries compiled before (Moein, 1998).

The corpus of the *Dehkhoda Dictionary* was larger and more diverse than all the corpora used by previous and later dictionaries. The components of the corpus could be categorized as follows:

1. Most of Persian and Arabic dictionaries compiled to that date
2. Persian texts in verse; for the first time, Dehkhoda read Ferdowsi’s *Shahnameh* and other collections of poems, extracting words and evidence from them
3. Persian texts in prose
4- Widely-used Persian words; in many cases, Dehkhoda would approach people to collect words and their meanings.

5- Historical and geographical announcements related to Iran, Muslim countries, and some non-Muslim countries.

Besides, Dehkhoda has explained scientific, literary, and grammatical and philosophic and issues as well as viewpoints and beliefs of sects and ideologies thus adding aspects of an encyclopedia to his dictionary (Khatibi, 2007). The publication date of this dictionary (1945-1980) is an important period in the history of Persian lexicography since it has been used as the main corpus for all the dictionaries that were compiled later.

3.4.3.2 Corpus in the Moein Dictionary

The publication of the Moein Dictionary was another important turning point in the history of Persian lexicography, indicating great progress in different fields and opening a new path for later lexicographers. Some features of the Moein Dictionary include: reasonable order of entries, sub-entries, definitions, and examples; Latin transcription; scientific etymology; and separation of proper nouns from words. Domination of Moein over the legacy of Persian language, especially Persian dictionaries, and employment of modern lexicography methods has made his dictionary as one of the main reference books used by speakers of Persian language.

The elements of the corpus used in the Moein Dictionary could be categorized as follows:

1) Persian texts from the beginning until the author's lifetime, including over 343 books and pamphlets.
2) Persian, Arabic and European dictionaries.
3) Arabic, Turkish and European encyclopedias such as *Mu'jam al-Buldan* by Yaqut, *Qamus al-A'alam* (Turkish), *Kashf al-ʿunūn* by Haji Khalifa *Atlas of Iranian Geography, Encyclopedia of Iranian Geography, Encyclopedia Britannica, Brockhaus Encyclopedia* Grand Larousse Encyclopedia, and *the Dehkhoda Dictionary*, which has been classified as an encyclopedia by Moein.

4) Words and names approved by the Persian Language & Literature House of Culture

5) Latest works on the etymology of Persian words by Iranologists, including *Grundriss der iranischen Philologie* by Wilhelm Geiger and Ernst Kuhn, *Grundriss der neupersischen Etymologie* by Paul Horn, and *Persian Studies* by Hübschmann

6) Notes taken from daily conversations and interviews with authors, poets, critics, and artists in different fields

7) Pictures, tables and maps (Khatibi, 2007).

### 3.4.3.3 Corpus in the Sokhan Comprehensive Dictionary

*The Sokhan Comprehensive Dictionary* is another important dictionary that has contributed to evolution in Persian lexicography. It was published in the 8 volumes in 2002 under the administration of Dr. Hasan Anvari. The dictionary consisted of two parts of old words and contemporary words. Introduction of the dictionary was titled as "word selection" briefly mentioned about pointed to the corpus used during its compilation process. Part of the introduction reads: "the linguistic corpus on which our dictionary has been based includes about two million examples of word usage mostly extracted from about four hundred old and modern texts. Our main focus has been on the current Persian. Our use of about one hundred and fifty contemporary short stories and novels as well as other sources in current Persian is indicative of such a focus. As
for Old Persian, we should admit that our resources have been limited compared to the wide range of old texts “(Anvari, 2002).

The failure to use sufficient older Persian texts is the main linguistic setback of this dictionary. Although the bibliography and content of the dictionary suggest that the most important Persian texts have been among its references, it has seemingly referred to other Persian dictionaries, especially the Dehkhoda Dictionary, or other dictionaries of old texts rather than Persian texts to select older words, phrases or usages as well as current phrases. The only difference is that it has used a scientific method to order entries and information related to each word and phrase. A review of the corpus of old texts used in the Sokhan Dictionary reveals that it shares a large number of references with the Dehkhoda Dictionary. Among the old translations and interpretations of the Quran, Tafsir al-Tabari is among the references of the Sokhan Dictionary (Khatibi, 2007).

3.4.4 Computational Corpus Linguistic

Today’s lexicography is closely interwoven with the computer. However, the traditional note-taking method has been used to compile general Persian dictionaries, including the Contemporary Persian Dictionary and the Sokhan Comprehensive Dictionary, in Iran over the past two decades.

The Persian Language Comprehensive Dictionary, presently under compilation by the lexicography department of the Persian Language &Literature Culture house, is the first Persian dictionary to be based on a large computational corpus of words taken from Persian texts.
3.4.4.1 Computational Corpora of Persian Language and Literature

Over the past two decades, a few Computational corpora has been developed in to Persian language. That has a great impact on the study of Persian Language and literature. Some of the important corpora’s include:

3.4.4.1.1 Dorj 3: The Compact Disc (CD) of the Mehr Argham Rayaneh Company

A new compact disc (CD) released by this company contains 178 Persian works in verse and prose from 101 authors and poets and included the texts from the oldest times to the current era. This CD is the first and best reference material for different types of research on Persian language and literature in terms of data size and application.

3.4.4.1.2 The Persian Language Database

This database has been created at the Institute for Humanities and Cultural Studies under the administration of Dr. MostafaA'asi. It is available at the institute's website. The database, designed to create a large collection of various corpora of Persian language, which includes selected literary, scientific, artistic and political Persian texts in written and spoken forms.

3.4.4.1.3 The Persian Language Corpus of the Agah Publications

This corpus was prepared under the supervision of Dr. Mohammad Reza Bateni at the Agah Institute during 1997-1999. It contains over 20 million words. The project had adopted a stylistic and sociologic approach toward Persian. Various usages of words from the current Persian language have been included in this corpus based on their level of importance. The corpus includes various texts such as stories, novels, political, philosophic, social and scientific texts, news items, articles, newspaper reports, contemporary magazines, classified advertisements, texts related to different aspects of daily life, and school textbooks.
Some features of this corpus are as follows:

1- This corpus enables lexicographers to access each word and the whole examples and usages related to it within the framework of lexicography and corpus linguistics standards with the help of a keyword tool.

2- It enables lexicographers to put words in alphabetical order based on their previous and next words.

3- It contains different formats and images of verbal contexts for word usage.

4- It is a very effective tool for recording lexical homophony.

5- It accurately determines the frequency of each word since texts have been added to it constantly, enabling lexicographers to select entries depending on the purpose and extent of their dictionaries (Khatibi, 2007).

3.4.4.1.4 Database of Persian language and Literature Texts

This project was approved at the Persian Promotion Council under the supervision of Dr. Mostafa Musavi in 2001. However, it started one year later. Based on the project, all Persian language books authored in different fields after 1921 were added to the database. The required software was designed and tested, and Persian texts authored during the ninth century and the first half of the tenth century was also added to it in the first phase of the project. However the project was stopped in 2003.

3.4.4.1.5 The Computational Corpus of the Persian Language &Literature House of Culture

The Persian Language &Literature House of Culture has prepared two corpora:

1- GanjVajeh (treasury of words), a corpus of Persian-Persian, English-Persian-English, German-Persian-German, French-Persian-French, Italian-Persian-Italian, and Spanish-Persian-Spanish dictionaries.
This project, carried out by the word selection department, included words and their meanings found in 375 dictionaries in different fields such as the humanities, medical science, agriculture, engineering, etc. The size of this corpus has reached to 6 million words so far. It also covers over 3000 words found in the glossaries of scientific books. The project is still underway.

2- The corpus of the lexical units of Persian texts,

This computational corpus of lexicography currently contains the lexical units of about 800 titles of Persian books (a total of over 100 volumes) and 7 million records. If each example contains at least 10 lexical units, the whole corpus will have about 70 million lexical units and examples. The corpus's software program has been designed with Foxbro in the user interface. (Khatibi, 2007).

The system will work as follows:

If one type a lexical unit the in the search field and the relevant sub-entries, examples, references and authorship date are shown immediately. If all the sub-entries and examples are required, click on the sub-entry count to observe the frequency of the lexical unit. Then, click on "export to Excel" option to access all the examples in chronological order. Another capability of this program is that it enables searching into the huge body of examples. Click on a button on the left of the example and a large number of examples will be provided in the chronological order in the Excel. Prefixes, infixes, and suffices could be easily searched within the text to examine their behavior in words and contexts.

3.5 Selection of Entries in Monolingual General Dictionaries

3.5.1 Definition of Entry

Entry is the point of access to information inside the dictionary. Entry can be defined as a series of words and phrases at the beginning of which a headword (the
dictionary's lexical and fundamental units) is placed. The headword is followed by information about it. It can be consisted of two main parts: the lexical unit and the word definition, which is a predicate that provides information about the lexical unit (Sharifi & Fakhimzadeh, 2007).

Entries may be divided into two categories, namely main and secondary entries. The main entry includes the full explanation and references related to the headword. The secondary entry is a derivation of the headword or a phrasal lexical unit that appears within the entry or at the end of it. It contains a usage label or other information about conjugation and examples. Homophones, antonyms, etc; are examples of secondary entry.

Lexical units serving as the subject of the main entry are called "headword" and words and phrases appearing in the secondary entry are called "sub-entry". Reduced entry is another type of entry which is consisted of the derivations of the headword that are not different from the headword in terms of meaning. Reduced entries appear under the related headword and provide no definition. Reduced headwords might be assigned as headwords in comprehensive dictionaries. The collection of a headword and its reduced sub-entries is called "nest".

Zero entry is still another type of entry that offers no information about the head word. In this type of entry, the headword is a word that has no usage by its own, but its phrasal forms are used. Such phrasal forms are defined as "zero entry". For example, the Persian term "Fart" (severity) is used in the phrase "As Fart-e" (from the severity of) and could be categorized as a zero entry depending on the lexicographer's decision (Sharifi & Fakhimzadeh, 2007).
Out of all the lexical units which exist in a language, only simple words and some complex and compound words can be selected as headwords. Frequency and enjoying vague or implicit meaning are the main requirements for complex and compound words to be chosen as a headword. The size and number of entries of a dictionary play a role in the process of selection. Issues related to the scope of time, geography, sociology, and dialect are also some on the aspects should be taken into consideration prior to process of the selection.

Recording the words of the standard language is the dominant tradition in lexicography. As for the Persian language, in addition to contemporary written words, words from the language spoken in the capital city (Tehrani Persian) which are found in newspapers, radio and TV shows are also used as entry in dictionaries. Today, extracting words from texts is done with the help of the computer and through the use of computational corpus. The main advantage of the computational corpus is that it offers more complete and correct information about the word and determines its frequency precisely.

3.5.2 The Role of Frequency in the Entry Selection Process

Frequency plays a very important role in the entry selection process of monolingual dictionaries. Recording a linguistic form becomes necessary when its frequency increases. Sometimes, it is also necessary to record a linguistic form even though its frequency is low, as and when a linguistic form is a reference word and has an important conception for speakers of the language.

Determining the headwords is one of the primary steps in the process of compiling a monolingual dictionary. Utilization of texts determined by the lexicographer and selection of words existing in those texts through linguistic initiation
has so far been the most common method of entry selection for monolingual general
dictionaries. When this method is used, it is possible that non-transparent common
linguistic forms with high frequency are not recorded while very transparent
complicated linguistic forms are recorded. Such a possibility will be higher if the level
of morphological productivity is not taken into consideration. Some theorists such as
Bain, Renov, and Dalton believe that it is improper to fill the dictionary with
transparent words that are easily understood with the help of morphological rules.

As the level of morphological productivity increases, the percentage of new
words not recorded in the dictionary will increase. It is noteworthy that most new words
are among single-frequency words. That is why an increase in the number of single-
frequency complicated words is attributed to its morphological productivity. Therefore,
increase in the reproducibility of the word-coining process should lead to a rise in the
percentage of single-frequency words coined during that process which is not recorded
in the dictionary. This principle could be a proper criterion for headword selection in
monolingual dictionaries (Abbasi, 2007).

Experts from scientific field should gather generalized scientific words; extract
scientific words that have entered the general language through searching into scientific
books, textbooks of technical schools, scientific journals, and glossaries. For example,
the term “Pathology” (pathology) observed on the signboard of laboratories and the
term “post-modern” that is widely used in the literature as well as daily radio and TV
conversations should be selected as headword in general dictionaries.

Many terms used in scientific fields have not been compiled yet. The same story
is true for jargons used by people in different walks of life. For example, market
traders, blacksmiths, shoemakers, public transport drivers, mechanics, etc. use of many
terms and expressions that is used in daily conversations, but has not been recorded by
dictionaries yet. Compiling these words and expressions requires that lexicographers
conduct field work and refer to people active in the above mentioned fields. Getting
access to many words that are used in daily conversations is only possible through
recording people's daily conversations or radio and TV programs. Among these words
are: naghsh-e tambr (postage stamp design), "tuduzi" (upholstery), and "safhekelaj"
(clutch plate). Oral texts should also be examined if such words include oral forms as
well (Sharifi, Fakhamzadeh, 2007).

3.5.3 Lexical Units (Non-Simple Words) and Entry Selection

Various types of non-simple words that could be selected as headword are as follows:

1- Compound words, including grammatical compound words such as "alive bar
   in" (in addition) and "ba vojude in" (in spite of this), and lexical compound
   words such as "barf pak kon" (windscreen wiper), "juje kabab" (barbecued
   chicken), "goje farangi" (tomato), and "tare ankabut" (spider web).

2- Complex words including one lexical morpheme plus one or more than one
   grammatical morpheme, such as "mardi" (manhood), "gereftari" (preoccupation), "kargar" (worker), and "danesh" (knowledge).

3- Complex compound words including several lexical morphemes plus one or
   more than one grammatical morpheme, such as "najavanmardi" (cowardice).

4- Repetitive words such as "sootsootak" (whistle) and "jighjighu" (screaming).

Zgusta (1971) suggests several criteria to identify multi-word lexical units:

1- Replacement-: usually, no part of a multi-word lexical unit could be replaced
   with another word.
2- It is impossible to add an element to a multi-word lexical unit. For example, it is not possible to say "bazare kheili siah" (the very black market) instead of "bazar siah" (the black market).

3- The meaning of a multi-word lexical unit does not come from the meaning of its components, because some components do not have their original meaning. For example, "sib zamini" (potato) is not a type of apple, "takhtesiah" (blackboard) is not necessarily black, and "dast" (hand) does not have its original meaning in the phrase "azdastdadan" (to lose).

4- A multi-word lexical unit might have a single-word synonym in the same language. For example, "koshtan" (to kill) is a synonym for "az dame tigh gozaran" (to massacre).

5- Sometimes, one synonym is found in a foreign language for a multi-word lexical unit. For example, the English equivalent of "mize tahrir" is "desk".

The above mentioned criteria are not true for fixed compound words. Some of these compounds do not have a role similar to that of a word based on paradigmatic and syntagmatic relations; for example, proverbs, quotation, fables, etc. In general, compounds in Persian language could be categorized as follows:

1- Compounds with "e" vowel (adjective compounds and ezafe compounds (a Persian linguistic construct)): The "e" vowel is omitted in some compounds; for example, "cheraghghoveh" (torch), "mize tahrir" (desk), and "takhtesiah" (blackboard).

2- Compounds with the word "va" (and) serving as their conjunction: for example, "gasht va gozar" (excursion) and "kamvabish" (relatively). Such compounds are chosen as headwords.
3- Compounds starting with a proposition: Some of these compounds such as "az kar oftadeh" (disused) are adjectives. Some of them such as "darbareye" (about) and "be vasileye" (by) are compound propositions, and some others such as "az anja ke" (since) are compound conjunctions (Sharifi, Fakhamzadeh, 2007).

3.5.4 Various Types of Dictionary Entries

The dictionary entry of words can be divided into several main groups, including nouns, verbs and adjectives, and several secondary groups, including pronouns, propositions, conjunctions and affixes. The dictionary entries can include expressions, abbreviations, brands, specialized words and phrases, loanwords, dialectic words, new words, etc. in addition to the above-mentioned groups of words. The general trend of dictionaries of Persian as follows:

1- Nouns: A noun makes up the majority of words of both animate nouns and inanimate, nouns are generally divided into singular nouns and plural nouns. Singular nouns such as "derakht" (tree), "khaneh" (house), "danesh" (knowledge), etc. are usually chosen as headword. Regular plural nouns are not usually recorded in the dictionary, while irregular plurals such as Arabic broken plurals and plurals ending with "at" as well as some plural nouns such as "mardom" (people), "galleh" (flock), and "khanevadeh" (family) are chosen as headwords. Gerunds and compound nouns such as "peydayesh" (emergence), "zibayi" (beauty), "amad o shod" (traffic) and "dardesar" (trouble) are also chosen as headwords.

1- Proper nouns are not usually recorded in general dictionaries, except for those proper nouns that have a special meaning or their compound and complex forms are used in the language. For example, "Isa" (Jesus), may not be recorded in the dictionary, but "Isavi" (Christian) is better to be recorded. Also, the names of
holy books, languages, tribes, places and individuals are recorded depending on different criteria (Sharifi, Fakhamzadeh, 1386).

2- Verbs: In dictionaries, a verb is recorded as a base from which other forms of the verb are produced. Various forms of the verb are not included in the dictionary, because native speakers can make them up through their lingual intuition.

Simple gerunds such as "khordan" (to eat), "didan (to see), "amadan" (to come), "keshidan" (to drag), etc are chosen as headwords. There are some special gerunds in Persian which turn into compound verbs when a linguistic element is added to them. Examples of these gerunds are: "zadan" (to beat, to strike) in "sohanzadan" (to file), "kardan" (to do, to make) in "tahdidkardan" (to threaten), and "khordan" (to eat, to consume) in "kotakkhordan" (to be beaten). These gerunds are known as "light verbs" in older grammar books (Ahmadi&Anvari, 1994). A new type of verb, called "prefixed verb", is also produced by adding some prefixes to simple gerunds. "bardashtan" (to remove), "faragereftan" (to learn) and "darraftan" (to escape) are examples of prefixed verbs.

Bone mazi" (past root) and "bone mozare" (present root) of simple gerunds are chosen as headword; for example, "khor" (eat) and "khord" (ate) from the gerund "khordan" (to eat) (Sharifi and Fakhamzadeh, 2007).

3- Adjectives: All simple adjectives are chosen as headwords. Compound adjectives are also selected as headwords if they enjoy high frequencies or opaque meanings; for example, "ghulpeikar" (gigantic), "tondro" (fast), "khoshmoshreb" (good-humored).

4- Adverbs: Simple adverbs are usually chosen as headword, while compound adverbs usually appear as sub-entries. For example, adverbs such as
"mehrabanane" (kindly) and "delirane" (bravely) which are coined by adding the suffix "ane" to an adjective and have transparent meanings appear as sub-entries under the relevant adjective.

Compound adverbs, such as "be tondi" (fast), "be deghat" (carefully) and "be arami" (slowly), which are consisted of a proposition and a noun and have transparent meanings, appear in bold letters under the relevant proposition.

5- Pronoun: All dependent and independent pronouns in Persian, such as "mim" (M), "t" (T), "man" (I), "to" (you), "ou" (he/she), etc., as well as other pronouns such as "khod" (self), "khodam" (myself), "khodeshan" (themselves), "yekdigar" (each other), "an" (that), and "in" (this) are chosen as headword (Sharifi, Fakhamzadeh, 2007).

6- Prepositions: Simple prepositions such as "az (from), "ba" (with), and "ta" (until) are chosen as headwords, while compound prepositions such as "be vasileye" (by) appears as sub-entries. There are also some exceptions in this regard. For example, the preposition "dabareye" (about) is selected as a headword because of its high frequency.

7- Conjunctions: Simple conjunctions are chosen as headword, while compound conjunctions appear as sub-entry. Compound conjunctions such as "benabar in" (therefore) and "chenanche" (in case) are exceptionally chosen as headwords because of their high frequency.

8- Function words: Function words are words with no lexical meaning that serve to express grammatical relationships with previous words. Exclamations such as "ey" (O) and "ya" (O) or "ra" (particle suffixed to a noun or pronoun ) which serve as an indirect object are among function words. Function words are chosen as headwords.
9- Affixes: an affix is a morpheme that has no lexical meaning and is always attached to another word to change its grammatical relation or create a new meaning. Productive affixes attached to numerous words are chosen as headword; for example, the suffix "nak" (tainted with) in "khatarnak" (dangerous) and "howlnak" (horrible) and "dan" (-holder) in "namakdan" (salt shaker) and "kahdan" (straw storage), or the prefixes "ba" (with), "bi" (without) and "na" (without) in "bahowsele" (patient), "bisalighe" (of bad tastes) and "najur" (ill-matched).

Unproductive affixes or the ones that are attached only to a limited number of words are not chosen as headword. Rather, the whole compound word is chosen as headword; for example, the suffix “stan” in “tabestan” (summer) and “zemestan” (winter) which is indicative of time. Words serving as verb root that are used to create compound words and have a role similar to that of affixes are also chosen as headwords; for example, “afshan” (-sprayer) in “zarafshan” (gold sprayer), “koob” (hit) in “ruyekoob” (person in charge of fixing upholstery), “saz” (-maker) in “kelidsaz” (key-maker), and “sakht” (-built) in “khoshsakht” (well-built). The importance of these words lies in the fact that they provide an unlimited capacity to coin new compound words in Persian.

10- Quasi-sentences: A quasi-sentence is a group of words that reflect the emotional states of the speaker such as admiration, pain, joy, etc. Words such as “ah” (oh), “afarin” (well done), and “ajab” (weird), and expressions such as “barikalla” (may God bless you) and “sobhanalla’ (glory to God) are examples of quasi-sentence.
The quasi-sentences usually serve as adverbs; for example, “afsus” (alas). Quasi-sentences are chosen as headwords for sure, although there are different viewpoints in this regard (Sharifi&Fakhamzadeh, 2007).

11- Expressions: Rubinchik(1992) believes that frequency of usage in the current language is not the only criteria for recording fixed expressions that could not be divided lexically, and that sometimes one or more than one component of such expressions lose their original meaning and convey an implicit meaning, making it necessary to record them. By identifying the pivot word of each expression, the expression could be fitted into an entry that has the pivot word as headword. “dastderazi” (aggression), “male mard om khori” (stealing from people) and “se kar kardan” (to blunder) are examples of expression.

12- Adages: There are still other groups of fixed compounds that are regarded as a single unit, but do not serve as a word in the sentence and are rather regarded as complete sentences. Proverbs, quotations, and maxims attributed to real or imaginary people that contain moral advice and have metaphorical meanings or verbal ornaments are among such groups of words. If an expression is assumed as a coherent chain of words, its entire meaning will come from the whole expression rather than the meaning of each of its components. An adage could also be categorized as an expression if, it offers a different meaning after one component of its replaced or omitted or one component is added to it. Some examples of adages are as follows: “hark e bamash bish barfsh bishtar” (a big head has a big ache), “kamgooy o gozide gooy chun dor” (few words are best), and “az mast ke bar mast” (accusing the times is but excusing ourselves).
There are different viewpoints as to recording adages in general dictionaries. Some scholars believe that adages should be recorded in a specialised dictionary. Zgusta (1971) believes that adages should also be recorded in large dictionaries such as comprehensive dictionaries. Adages should be included as sub-entries if the dictionary targets non-native speakers.

13- Abbreviations: An abbreviation is coined through juxtaposing the first letters of the words forming the name of a phenomenon in order to shorten the phrase in question. Abbreviations are used in daily language and some of them take the shape and spelling of a word; for example, “homa” (fabulous bird of good omen) is the abbreviation of “havapeimyi melli Iran” (the Iran Air). Some abbreviations are pronounced word by word; for example, BBC, which is the abbreviation of British Broadcasting Corporation. All abbreviations appear as headwords in dictionaries, because they function like word (Sharifi, Fakhamzadeh, 2007).

14- Brands: Some brands (usually the names of manufacturing companies) are used to refer to the products of the company eventually, for example, “Pepsi”, “Rika”, and “Xerox”. Such brands should be chosen as headwords.

15- Non-organic compounds: Non-organic compounds are compounds consisting of separate parts of a word or morphemes from several words. The word “razmayesh” (military exercise) is an example of such compounds. These compounds are chosen as headword.

16- Loanwords: Loanwords usually come from foreign languages. Recording loanwords is one of the main tasks of lexicographers. Some Arabic loanwords used in Persian have undergone so much phonetic and semantic change in Persian language that they have lost their original meanings
and replaced their Persian equivalents. Some of these loanwords include “en’ekas” (reflection), “beinolmelali” (international), “nokha” (marrow), and “vaba” (cholera). It is necessary to record such words in the dictionary. Also, some Arabic expressions such as “ensha’alla” (God willing) and “rahmatollahalayh” (peace be upon him) that have become clichés in Persian are also recorded.

Many European loanwords such as “telefon” (telephone), “radio” (radio), “televizion” (television), “taksi” (taxi), “mikrob’ (microbe), and “virus” (virus) are also recorded in general dictionaries (Sharifi& Fakhamzadeh, 2007).

17- Dialectic words: Speakers of various dialects in different parts of the country exchange their cultural values through cultural, economic and familial relations, thus loaning words to each other. The standard Persian is filled with such words which have been recorded and are used in daily conversations. The words "gabbeh" (rug) from Luri dialect "varza" (ox) from Mazandarani, and Gilak idialects are among loanwords. Dialectic words are recorded in dictionaries.

18- Specialized (technical) words: Experts engaged in different fields as well as lexicographers of general dictionaries generally identify and record specialized words. Specialized words such as "lizing" (leasing), "disk" (disc), and "patolojy" (pathology) which are used in the general language should be recorded as headwords.

19- New words: New words enter vocabulary in two ways: 1. through natural rules; for example, "anteneboshghabi" (dish antenna), and 2. through an organized word-coinning process; for example, "pasandaz" (saving) and "daneshgah" (university), which have been coined by the Persian Language & Literature House of Culture.
New words are not recorded in dictionaries as long as they are not used widely. For example, when the words "charkhbal" and "balgard" which have been offered as Persian equivalents for "helicopter" could be recorded in dictionaries when it turns into a widely-used word over a period of time (Sharifi, Fakhamzadeh, 2007).

3.5.5 The Need of Entry

In general Persian dictionaries, all words are recorded based on their Persian spelling and their pronunciation in the standard language. Sometimes, some words are used as headwords more than once. This is true for similar words such as the ones with similar spellings and different pronunciations. These words are chosen as headwords more than once depending on the number of their pronunciations. Examples of these words are "dar" (door) and "dor" (pearl), and "karam" (generosity), "kerm" (worm), and "kerem" (cream).

Words with similar spellings and pronunciations but different origins or meanings are recorded separately; for example, "masaleh" which means interests and "masalah" which means construction of materials.

3.6 Etymology in Persian Dictionaries

The word etymology is among the issues addressed during the compilation of monolingual dictionaries in different languages. Word etymology has gained a special status thanks to the emergence of modern approaches in language studies especially linguistics since the eighteenth century. It becomes a part of any entry in the lexicon which includes root word, the history of words and its development and various changes.
From the viewpoint of linguistics, accomplishments in etymology include knowledge about the history of each word or the spoken form of it, which are useful for various general and specialized dictionaries. Monolingual dictionaries usually provide etymology sections as well. Research on the etymology of words should be associated with moderation always. It means that the historical background of the word in terms of phonetic and phonological changes should be studied rather than its roots in pre-history eras. Also, the etymology section should not be limited to the mere semantic changes of the word and should not ignore the origin of the word (Sarkarati, 1998).

These are the viewpoints one has to keep in mind during the development of dictionaries or any selected assignments. It is worth mention here the compilation of Dravidian Etymological Dictionary by T. Burrow and M.B. Emeneau. This dictionary is a compilation of words and meanings from various sources on the Dravidian Languages. "This dictionary does not contain proto Dravidian (Pdr) reconstructions. This would have been useful, but it was not thought that the considerable extra expenditure of time that would have been necessarily to prepare them was warranted in the present stage of Dravidian studies. The object of the dictionary is to provide material for such studies not to record results which at the moment could be little more than superficial and non definitive judgments” said the authors.

When you set out to make a monolingual dictionary, you have to decide whether you want to include etymology. The answer to that question depends on the aim of the dictionary. If the aim is to serve as a learning dictionary, etymological information is not needed. However, if the dictionary has a historical component, for instance because it intends to describe the vocabulary of the past fifty or hundred years (as the ‘Grote Van Dale’ does in the case of Dutch), the inclusion of etymology is the logical consequence of that choice.
Some of the etymological dictionaries should be categorized as specialized dictionaries since they include the complete etymology of each entry. Etymological information could also be included in general dictionaries as part of each headword. *The Moein Dictionary* and *the Sokhan Comprehensive Dictionary* are among Persian dictionaries with etymology sections. Etymological information could be found in the following types of dictionaries:

1- Etymological dictionaries


3- Dictionaries of Persian in the Islamic-era (such as *the Historical Dictionary of Persian Language* by Natel Khanlari et al., 1996).

4- Monolingual general Persian dictionaries (such as *the Moein Dictionary*).

5- Dictionaries of Persian loanwords (such as *Turkish and Mogul Elements in Persian Language* by Doerfer, 1963).

6- Specialized Persian dictionaries (such as *the Dictionary of Idiomatic Expressions during Qajar: Army and Police* by Modarresi et al., 2001).

Etymology has turned into an inseparable part of English, French, and Russian dictionaries for years now. However, it is not still too popular among Persian dictionaries. Recording etymological information in monolingual dictionaries is not only a lexicographical requirement, but also makes such dictionaries suitable for non-lexicographical purposes as well. It is especially true for Persian language since it lacks comprehensive etymological dictionary presently.
Words are in fact the most elusive part of each language. They appear and disappear, and are always exposed to changes. Most language changes occur to words. Many Persian words have undergone such changes over long periods of time, making it difficult or impossible to identify their origins and even the date of their appearance. According to Landau (1991), the loss of the original meaning of a word due to distancing from its etymological meanings will push that word towards folk etymology. It is even difficult sometimes to determine whether a word is a loan word or an original word. For example, the word "khashen" means dark in Middle Persian, while in Arabic "khashen" means rough and ill-natured. It is not clear if the Middle Persian world has undergone semantic change or the two words are not related to each other at all. Another example is the Persian word "kheng" which has two original meanings: 1- white horse, and 2- fool. It is not clear if the word has multiple meanings or each meaning belongs to a different entry. Such words are numerous in Persian. In Persian, ambiguities in etymology are mainly related to colloquial and vulgar words as well as words that have not been recorded in older Persian dictionaries (such as Borhan-e Ghate and Loghat-e Fors dictionaries) (Bakhtiari, 2009).

In spite of all difficulties and confusions in the etymology of colloquial and vulgar Persian words, scholars repetitively address the etymologies of original words such as "khordan" (to eat), "amadan" (to come), "didan" (to see), "kuh" (mountain), "darya" (see), "khorshid" (sun), "aseman" (sky), and alike, which could be traced to Indo-European languages, or continue to forge etymologies for words that either lack originality in terms of phonetics and semantics or are not ambiguous at all in terms of meaning and reason of appearance (Zakeri, 2002).

Getting far from the origin is not limited to ancient words and is also observed among contemporary words or phrases with ambitious origins or reasons of appearance.
Examples of such words are "rakab zadan/khordan" (to deceive/ to be deceived), "do darkardan/shodan" (to hoax/ to be hoaxed), "toop" (excellent), and "terip" (gesture). It would be better for dictionaries that record such words to leave research on the etymology, semantic evolution and appearance date of these words to future lexicographers. Lexicographers have their own reasons to include etymological information in their dictionaries. Some of these reasons are as follows:

1- To separate entries accurately

2- To explain some linguistic phenomena such as the existence of different phonetic forms for words such as "toshak/doshak" (mattress) and "dokme/tokme" (button). Etymology tells us, for example, that these different phonetic forms have come from Turkish dialects.

3- Linguistic planning and language teaching

4- Familiarity with the semantic evolution of words

The multilingual dictionaries normally do not have etymology sections. The same story is true for the dictionaries for children. Etymology essentially targets adults and native speakers of the language. However, glossaries usually have etymology sections, especially glossaries of ancient sciences such as medicine as to Medical terms are of Greek and Latin origins mostly (Bakhtiari, 2009).

3.6.1 Etymology in Etymological Dictionaries

An etymological dictionary is a dictionary that mainly seeks to offer information about the origin of words to experts or non-expert users. There are some differences between etymological dictionaries and other types of dictionaries. For example, etymological information is considered secondary in the compiling of general dictionaries as its function is very limited. While in etymological dictionaries the whole
entry is dedicated to phonological and semantic information of historical nature, serving
goals mentioned in the definition of etymology.

### 3.6.2 Methods of Providing Information in Etymological Information

Various criteria have been suggested for compiling Persian etymological
dictionaries some of which are as follows:

1. To determine chronological history: it means to which era the origin of the
   word dates back; for example, whether (a) to the period of time from the
   Ancient Iran era to the Modern Iran era, (b) to the Indo-Iranian era or Indo-
   European era.

2. To determine the scope of the language: it determines that to what extent the
   original form of a word should be searched or reconstructed and to what extent
   it should be compared with other languages. If the relevant research is too
   broad, all various forms of the word in other languages or dialects should be
   mentioned.

3. To determine direction: to determine whether to start from the current forms of
   the word and move back to its older and hypothetic forms or to do vice versa.

4. To select linguistic corpus: to determine based on which corpus etymological
   research should be conducted.

Historical dictionaries cover the long history of recorded words and aim to trace
the development of words from their original usage in literary and non-literary texts to
the present day in chronological order, in order to determine:

1. Existence of each word in different eras, places, tribes

2. Change in word's meaning, usage, form and pronunciation.

3. Word's usage in phrases
4- Word's origin and derivations
5- Word's morphological and syntactic features
6- Word's stylistic and statistical features

In an introduction to the first volume of *the Oxford Dictionary* (1888), Sir James Murray writes: this dictionary seeks: 1- to show how each single word has entered English language, how its form and meaning have transformed over time, how have been its style and usages, and what have been its new usages, 2- to show these information through the use of chronologically ordered evidence, and 3- to offer etymology to identify each word" (Bakhtiani, 2009).

The importance of etymological information and the need for including such information in a historical dictionary have been highlighted in a statement by Landau (1991): "the importance of etymology in historical dictionaries is beyond imagination. Mentioning the nominal and semantic evolution of a word in a scope of time is the main goal that a historical dictionary follows. It corresponds with etymology in many aspects, although it is not exactly equivalent to etymology.

3.6.3 Etymology in Dictionaries of Loan words

Dictionaries of loan words are among the dictionaries in which the origin of words is mentioned. Etymological explanations are more detailed in some dictionaries of loanwords. Examples of dictionaries of loanwords in Persian and other Iranian languages include *the Dictionary of Ancient Iranian* by Bartholomae, *the Etymological Dictionary* by Horn and a criticism on it by Huisman, *the Concise Pahlavi Dictionary* by MacKenzie, *the Pahlavi Dictionary* by Nyberg, *the Soghdian Dictionary* by Badrozzaman Qarib, etc. These dictionaries provide etymological information depending on their procedures and methods. However, a dictionary of loanwords is a
dictionary which records and examines loanwords. As for ancient Iranian languages, the *Altiranisches Sprachgut der Nebenuberlieferungen* by Walther Hinz (1975) addresses these languages. Although this book does not address the loanwords of ancient Iranian languages, it principally reflects loanwords from Iranian dialects such as the Median dialect.

A number of studies, both scholarly and non-scholarly, have also been conducted on Arabic loanwords existing in Persian and Persian loanwords found in Arabic. Among scholarly works is Arthur Jeffery's *The Foreign Vocabulary of the Qur'an*, which is a research on the etymology of some words and names used in the Qur’an and shows nominal and semantic evolution in these words and names from their original forms to their Qur’anic forms (Bakhtiari, 200). These words have been dubbed "arabicized words" since old times, namely words from foreign languages that have been added to Arabic language. Persians words make up part of these words. Some of the Persian words include "ferdows" (heaven), "serat" (path), "zoor" (vigor), "rowze" (commemoration of religious figures) and "rezgh" (daily bread).

*Al-mo'arrebva al-dakhil* authored by Sa'diZonavi (2004) is among the books that have scientifically examined arabicized words, especially arabicized Persian words. Arabicized Persian words make up a major part of the words recorded by this dictionary. Ambiguities are also observed in the book, showing that its author has gone to extremes in some cases. Some of the Persian words recorded by this dictionary include: "abdast" (pious), "abzan" (bathtub), "astane" (threshold), "salar" (chief), "shotorban" (camel-driver), "asgar" (army), "karkhane" (factory), "kamandar" (archer), etc. The origin of the word has also been mentioned in some cases; for example, "ajor" (brick), from Persian word "agor"/"agur", origin: Aramaic. As for works addressing Indian words in Persian language, *Indian Words in Persian* in Urdu (authored by
Abdolrashid, 1996) and articles by Sarkarani (1999) and Rezayee-Baghbidi (1996) could be mentioned.

A huge four-volume work by Professor Doerfer (1963) is the only scholarly and complete work on Turkish and Mogul words used in Persian. Most Turkish loanwords, apart from a small number of vulgar loanwords that have recently entered Persian, could be found in this dictionary. However, only a limited number of these words are unknown in terms of etymology (Bakhtiari, 2009).

3.6.4 Etymological Data in Persian-Persian Dictionaries

Around 28 general Persian dictionaries have been compiled in Iran over a 65-year period between 1924 and 1989. The Nafisi Dictionary was the first dictionary of this type, followed by the Dehkhoda Dictionary, the Nezam Dictionary and the Moein Dictionary. Several other dictionaries were also compiled between 1989 and 2003. The Sokhan Comprehensive Dictionary was among these dictionaries.

Among the authors of the above mentioned dictionaries, only two lexicographers have paid attention to the etymology of words: They are viz, Mohammad Daeeioleslam, who has compiled the Nezam Dictionary, and Mohammad Moein, who has authored the Moein Dictionary and edited the Borhan-e Ghate Dictionary. Moein has had access to more information, but the origins of some of the words especially loanwords in his dictionary remain unknown. Also, the volume of etymological information provided for various words by the above-mentioned lexicographers is not equal. Besides, the dialectology information has been mixed with etymological information. Same’i (1989) mentioned that these lexicographers have included what they have known rather than what they should have known.
3.6.4.1 Etymology in the Nezam Dictionary

Daeeiologieslam was the first lexicographer to author a Persian dictionary based on scientific and linguistic principles as he always gone in touch with Iranian cultural and linguistic institutes such as the Persian Language & Literature House of Culture (Same'i, 1989). However, the etymological information provided in his dictionary is poorer than the scientific structure of the whole work.

Daeeiologieslam's scholarly approach toward the two fields of non-poetical language and spoken language as well as his attention to differences between men and women are also interesting. He has innovated etymology in Persian-Persian general dictionaries, making sporadic efforts to provide etymological information based on knowledge and facilities existing at the time. However, Persian etymology would have been in better a condition today if his elementary method had been followed and improved by other lexicographers (Bakhtiai, 2003).

3.6.4.2 Etymology in the Moein Dictionary

*The Moein Dictionary* has seemingly been more successful in providing etymological information since its author has had access to more information. In the "etymology method" section of the dictionary, Moein(1997) has explained:

1) The origin of each word and its etymology in the original language has been mentioned in a bracket ([ ]). (page, 277)

2) In this way, we have tried to be concise. It means that we have only mentioned the Pahlavi form of the Iranian words, if the Pahlavi form has been available (except for a small number of cases where it has been necessary to mention the Avestan or Old Persian form of the word as well) or included the Sanskrit form of the word if the Pahlavi form has not been available. For Persian words with
no roots in ancient Iranian languages, words of the same root in Persian or
samples from Iranian dialects have been mentioned.

3) The origin of the word has not been mentioned for words that are originally
Persian, unless they have been combined with other words.

4) Explanations have been provided for abbreviated words, "mazid elayh" words
(words that do not lose their original meaning even when extra letters are added
to them) and "momal" words (words in which the letter "a" could be replaced
with the letter "e").

The etymological advantages of this dictionary include:

1) Mentioning the primary form of foreign words; for example, "azughe"
(provisions) (Turkish) {"azugh", "azogh"}

2) Providing transcription for old and middle Iranian words; for example, "bozorg"
(big) {"bazurg" in Pahlavi}

3) Mentioning of reconstructed forms; for example, "peida" (visible) {"ay" (old
Iranian)}

4) Providing etymological references; for example: "tutia" (vitriol), {arabicized} 
"tutia" (German)tuthia).

5) Clarifying information through the use of languages of the same family; for
example, "bagh" (garden): bay (Pahlavi), bay (Soghdian)

6) Mentioning the semantic evolution; for example, "achar" (screw-driver):
{Turkish, "achar", "achmagh": to open, to unfasten}

7) Avoiding to get overwhelmed with the etymology of vulgar terms

Bakhtiari (2003) rightly pointed out that; the Moein Dictionary is the first
Persian dictionary to provide such important etymological advantages. He further adds
that, it is surprising that etymology in Persian dictionaries has remained immature in
spite of significant progress in linguistics especially in historical linguistics and its success in discovering and reviewing Old Iranian, middle Iranian and non-Iranian texts. Even the compilation of dictionaries of words extracted from these texts, explanations on the origins of words and their relation with other Iranian and Indo-European languages, progress in Iranian dialectology, and compilation of huge works such as Nyberg's *Manual of Pahlavi*, Doerfer's *Turkish and Mogul Elements in Persian Language*, Bailey's *Dictionary of KhotanSaka* (1979), Badrozzaman Qarib's *Soghdian Dictionary*, and alike have not made any improvement on this in Iran. The reason may be Iranian lexicographers have not been linguists and have not had ties with linguists. They have been lexicographers engaged in etymology rather than being linguists engaged in lexicography and etymology as mentioned (Bakhtiari, 2003).

3.6.4.3 Etymology in the Sokhan Dictionary

*The Sokhan Dictionary* has paid more attention to etymology. It explains: "in our dictionary, etymological information is offered in a bracket [ ] immediately after providing transcription. Etymological information is not too detailed in that dictionary. Wherever no etymology bracket is seen, it means that the headword is of Persian origin.

Wherever the origin of the word is unknown, a question mark is placed in the etymology bracket. The origin of the word is not mentioned for adjectives attributed to proper nouns, such as "afrighayi" (African) and "italiayi" (Italian). In such cases, only the origin of words with Arabic structure such as "alavi" (descendant of Imam Ali) and "mostafavi" (descendant of Mostafa) is provided.

Wherever the headword is a loanword from another language, its original form is mentioned. The name of the foreign language from which the word has been taken appears in abbreviations; for example, "Ar." stands for "Arabic". Arabic words make up
the majority of foreign words existing in Persian. If a word is Arabic and is used in Persian with the same meaning, only the fact that it is Arabic is mentioned; for example, "aghel" (wise): {Ar.}. If an Arabic word has undergone phonetic or spelling changes in Persian, its Arabic form will be mentioned as well; for example: "an'am" (tip): {Ar., en'am}.

For loanwords from languages with Latin alphabets, such as French, English and German, the Latin spelling of the word is also mentioned; for example, "kuler" {Eng., cooler}. For arabicized words, the original language is often mentioned; for example, "esfanj" (sponge): {arabicized from Greek}. Wherever an Arabic word is plural, the singular form of the word is also mentioned; for example, "asaar" (works, effects) {Ar., Pl of "asar" (work, effect). Other types of information are also included in the etymology bracket depending on the word in question; for example, "ajan" (agent): {Fr., = azhan} (Anvari, 2002).