2. Statistical survey of the Data of Words (DW)

This chapter gives a statistical description of the Data of Words (DW). It is useful to get an overview of the DW, since the Data of Basic Syllables (DBS) is extracted from it. As was said before, it is the DBS that serves as the basis for the analysis of Persian syllables in this work.

First I shall deal with the etymology and structure of the words in the DW. Next, I shall look at frequencies of individual consonants and of different consonant groups. Then vowel frequencies and frequencies of different vowel groups will be presented. The data will be divided into different groups on the basis of the etymological origin of words, after which consonant and vowel frequencies will be given for these groups. Consonant and vowel frequencies will be compared between different etymological groups.

2.1. Etymology and structure of words in the DW

As mentioned in 1.5.3, the DW consists of 10175 words. In addition to etymologically Persian words, it also contains words borrowed from thirteen other languages:[32] Arabic, French, Turkish, Greek, English, Russian, Sanskrit, Hindi, Aramo-Syriac (Aramaic and Syriac), Latin, Italian and Chinese (Appendix 1). Some words in the DW are etymologically mixed, i.e. they have elements from two languages. There are six types of etymologically mixed words in the DW. The following list gives one example of each type:

---

32 I consulted the following dictionaries to determine the etymology of words: Steingas (n.d.), Nafisi (1976), Padeshah (1984), Razi (1987), Moshiri (1992) and Dehkhoda (1993).

(1) Arabic-Persian: /su-rät-gär/, 'portrait painter', 'sculptor', (Arabic /su-rät/ + Persian suffix /gär/)
(2) Persian-Arabic: /sa-de-loh/, 'naive', 'simpleton', (Persian /sa-de/ + Arabic /loh/)
(3) Turkish-Persian /sag-duš/, 'groomsman', (Turkish /saG/ + Persian / duš/)
(4) Persian-Turkish /pa-tuG/, 'haunt', 'place of rendezvous', (Persian /pa/ + Turkish /tuG/)
(5) French-Persian /film-bär-dar/, 'cinematographer', (French /film/ + Persian /bär-dar/)
(6) Sanskrit-Persian /nil-gun/, 'blue', 'cerulean', (Sanskrit /nil/ + Persian /gun/)

Altogether 39.30% of the words in the DW are etymologically Persian, while 59.74% are loanwords and less than 1% are etymologically mixed. The great majority of loanwords come from Arabic, their proportion being about half (52.91%) of the DW. French is the second biggest contributor of loanwords, with a proportion of 4.94% of the DW. The following list shows the percentages of major etymological sources of the DW. A more detailed version of this list is in Appendix 1.

	No. of words	%
Arabic	5384	52.91
Persian	3999	39.30
French	503	4.94
Turkish	105	1.03
Bilinguals	97	0.96
Other languages	87	0.87
Total	10175	100.00

All of the loanwords in Modern Persian have undergone nativisation processes of Persian.[33]

[33] Nativisation processes are phonological processes of Persian, filters through which all foreign words compulsorily pass. The following is an example list of changes Arabic consonants undergo during the persianisation processes:

The DW contains 24135 syllable tokens. The frequencies and percentages of the three syllable types are as follows:

Syllable type	Frequency	%
CV	11126	46.10
CVC	11874	49.20
CVCC	1135	4.69
Total	24135	100.00

We see that almost half of the syllable tokens are of the type CVC, and that CV is nearly as common, while the most complex syllable type CVCC only covers about 5% of the tokens. Etymologically, 7.42% of the syllable tokens in the DW are exclusively[34] of Persian origin, and 14.82% are exclusively of Arabic origin (see Appendix 4).

Table 2.1 shows different word types[35] in the DW, together with their frequencies and percentages. The number of syllables in a word ranges from one

(1) /ð̣/ (emphatic, dento-alveolar, fricative), /d̩/ (emphatic, stop), /ð/ (interdental, fricative) → /z/ (alveolar fricative)
(2) /s̩/ (emphatic, dento-alveolar, fricative), /ɵ/ (non-emphatic, dento-alveolar, fricative) → /s/ (alveolar fricative)
(3) / ħ/ (pharyngeal, fricative) → /h/ (glottal fricative)
(4) /ʕ/ (voiced, pharyngeal, fricative) → /ʔ/ (voiceless, glottal stop)
(5) /w/ (labial, semivowel) → /v/ (labiodental, fricative)
(6) /t̩/ (emphatic, dento-alveolar, stop) → /t/ (dental stop)

Degemination is one of the persianisation processes that Arabic words particularly undergo. It reduces sequences of syllable final identical consonants to one. For example: *sädd → säd, fänn → fän.* Two other very common nativisation processes of Persian are *prothesis* and *anaptyxis*. The process of prothesis operates in two ways. If a word and/or syllable starts with a vowel, the process inserts a glottal stop /ʔ/ before it. For example: *écran (French) → ʔekran*. Should the word and/or syllable begin with a consonantal cluster, and if the first member of the cluster is /s/ or /š/, the process of prothesis inserts a sequence of /ʔe/ before the cluster. For example: *ski (English) → ʔeski, škoda → ʔeškoda*. The process of anaptyxis influences words or syllables starting with a cluster other than /sC-/ and /šC-/, and splits the clusters by inserting anaptyctic vowels /o/, /i/ or /e/. For example: *plaque (French) → pelak, bronze (French) → boronz, cravate (French) → keravat, grammaire (French) → geramer*. Arabic words are not influenced by these processes because both Persian and Arabic syllabic structure does not permit syllable initial vowel or consonant clusters.

34 Here, by 'exclusively' I mean that similar syllables were not found in other contributing languages.

35 The term *word type* refers in this study to the number of syllables in a word.

to seven. The most common word type in Modern Persian is disyllabic, with a share of nearly half of the DW (47.43%), and the next common type is trisyllabic, comprising roughly a third (34.15%) of the DW. The third most frequent type is monosyllabic words, with a share of a little over ten percent (11.48%). Thus, the proportion of multisyllabic words in the data is nearly 90%. We see that the longer the multisyllabic word is, the lower its frequency is.

Table 2.1: *Frequencies and percentages of word types in the DW, in descending order*

Word types →	**2-syll.**	**3-syll.**	**1-syll.**	**4-syll.**	**5-syll.**	**6-syll.**	**7-syll.**	**Total**
Frequency	4825	3474	1170	651	43	11	1	10175
Percent	47.43	34.15	11.48	6.40	0.42	0.11	0.01	100.00

The most common word pattern[36] in the DW is CVC-CVC (see Appendix 2). This pattern comprises 17.68% (1799 words) of the words in the data. The next common pattern is also disyllabic, CV-CVC, with a proportion of 16.97% (1726 words). The ten most frequent multisyllabic word patterns are all either disyllabic or trisyllabic, and they contain only CV and CVC syllables. The top ten list of (all) word patterns contains two monosyllabic types: CVCC with a proportion of 6.31% (644 words) and CVC with a proportion of 4.93% (454 words). The third syllable type, CV, can also occur as a word, but rather infrequently (0.23%, 23 words).

Table 2.2 displays the frequencies of different syllable types in different positions (i.e. as the first syllable, second syllable, etc.) of a multisyllabic word. The numbers 1, 2, 3, etc. on the heading row refer to the seven possible positions where a syllable can occur in a word. The table shows that more than half of word-initial syllables, are of the CV type (5109 syllables, 56.74%), and clearly more than a third are of the CVC type (3757 syllables, 41.72%), but very few are of the CVCC type (139 syllables, 1.54%). The CVC type is clearly the most frequent as the second, third, or fourth syllable, with the CV syllable coming next. The CVCC syllable is the least frequent type in all positions of a

[36] The term *word pattern* refers here to the syllabic structure of the word. For example, the word /da-ʔe-rä-tol-mä-ʔa-ref/ has the pattern CV-CV-CV-CVC-CV-CV-CVC.

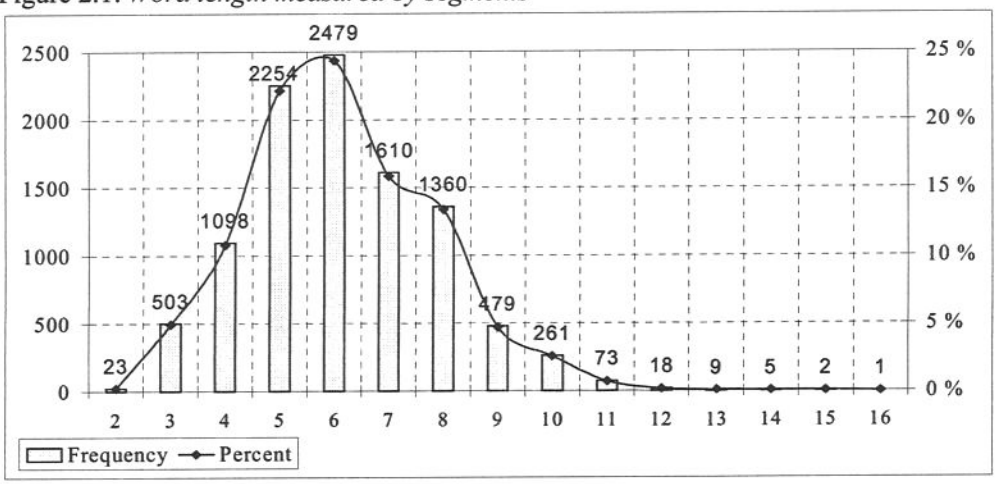
multisyllabic word. We saw above that the CVCC syllable has the highest frequency as a monosyllabic word.

Table 2.2: *Frequencies of syllable types in different positions of multi-syllable words*

Syllable patterns	1	2	3	4	5	6	7	Total
CV	5109	3990	1652	307	37	8	0	**11103**
CVC	3757	4768	2436	389	16	4	1	**11371**
CVCC	139	247	93	10	2	0	0	**491**
Total	**9005**	**9005**	**4181**	**706**	**55**	**12**	**1**	**22965**

In the previous paragraphs, the word length was measured by the number of syllables. We can also measure the length by the number of segments that a word contains. Figure 2.1 shows the frequencies and percentages of words of different length, measured by the number of segments. The figure shows that the word length in the data ranges from two to sixteen segments. The most common word length is six segments; about 25% (2479 words) of the words consist of six segments. This corresponds to the word patterns CVC-CVC, CV-CV-CV, CV-CVCC and CVCC-CV, the first two of which are very common. Next in order are words with five segments (2254 words, 22.16%), i.e. word patterns CV-CVC and CVC-CV. The length of the vast majority of the words in the data is between four and eight segments (8799 words, 84%), and practically all words are at most thirteen segments long.

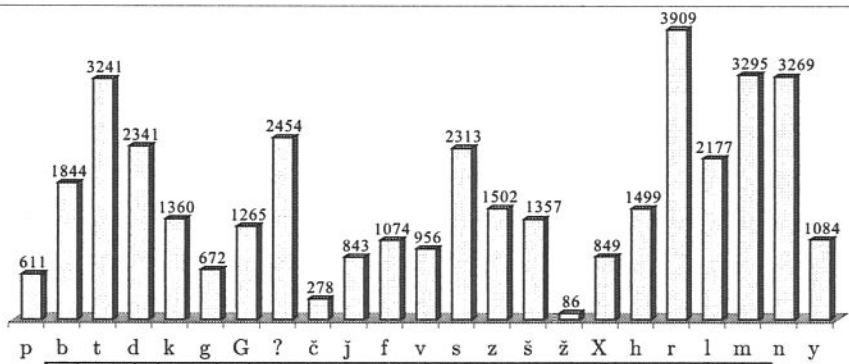
Figure 2.1: *Word length measured by segments*



2.2. Consonant frequencies

The DW contains 38279 consonant tokens. As Figure 2.2 shows, they are not evenly distributed over the 23 consonants; the most frequent consonant /r/ covers about 10% of the occurrences, while the least frequent consonant, /ž/, only takes 0.22%.

Figure 2.2: *Frequencies and percentages of consonants in the DW (DO = in descending order)*



p	1.60%	d	6.12%	G	3.31%	j	2.20%	s	6.04%	ž	0.22%	r	10.21%	n	8.54%
b	4.82%	k	3.55%	?	6.41%	f	2.81%	z	3.92%	χ	2.22%	l	5.69%	y	2.83%
t	8.47%	g	1.76%	č	0.73%	v	2.50%	š	3.55%	h	3.92%	m	8.61%		100.00%

DO: r>m>n>t>?>d>s>l>b>z>h>k>š>G>y>f>v>χ>j>g>p>č>ž

Of the consonant tokens, 64.13% are obstruents and 35.88% are sonorants. But even though sonorants as a group have a smaller proportion than obstruents, the frequencies of individual sonorants can be high. As the hierarchy in Figure 2.2 shows, the three most frequent consonants are all sonorants, i.e. /r/, /m/, /n/. These three together cover 27.36% of all consonant occurrences in the DW. On the other hand, some obstruents have very low frequencies, e.g. /ž/ and /č/ are remarkably rare. Since the class of obstruents has more than three times as many members (18 consonants) as the group of sonorants (5 consonants), the class frequency (i.e. the sum of the frequencies of all consonants in the group) at least partly reflects the size of the class. To eliminate the effect of the different class sizes, the average frequency per consonant in each class was calculated. The average frequencies per obstruent vs. sonorant consonant are given in Table 2.3.

We see that the average frequency per sonorant is about twice as high as the average frequency per obstruent consonant.

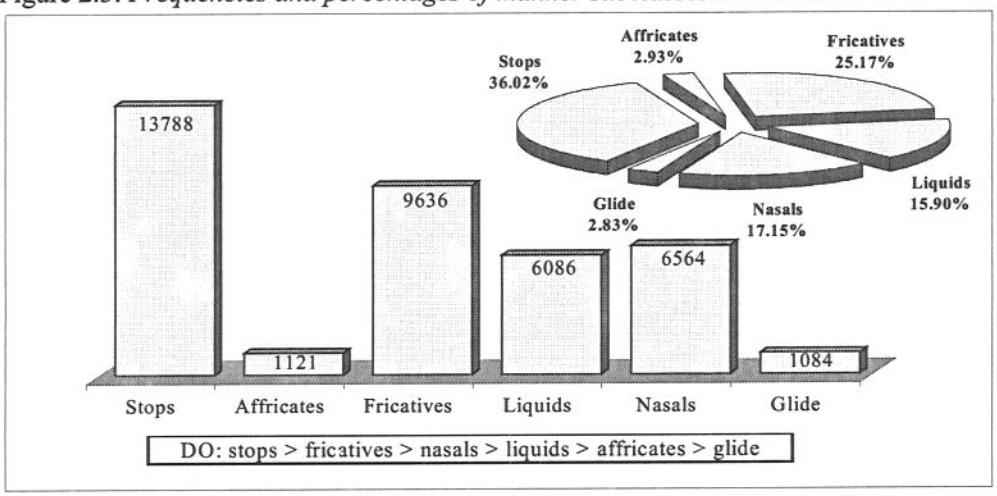
Table 2.3: *Average frequencies of manner classes in the DW*

Obstruents	Sonorants
1363.6	2746.8

Of obstruent tokens, 61.26% (15036) are voiceless and the rest, 38.74% (9509) are voiced. We discussed in 1.4 that most obstruents occur in pairs voiceless/voiced; those that do not are the voiceless /χ/, /h/ and /ʔ/ and the voiced /ɢ/. Thus, there are two more voiceless than voiced obstruents. But this alone does not account for the bigger proportion of voiceless obstruents. In most voicing pairs, the voiceless member is more frequent. The only exceptions to this are the affricate pair /č/ and /ǰ/ and labial stop pair /p/ and /b/.

With regard to the manners of articulation, stops are the most frequent manner subclass, as seen in Figure 2.3. They take more than a third (36.02%) of all consonant occurrences in the DW. Fricatives, the next biggest group, take a quarter (25.17%) of consonant tokens.

Figure 2.3: *Frequencies and percentages of manner subclasses in the DW*[37]



37 Note that in this figure and all other figures and tables that deal with the manner subclasses, the term *glide* refers to the phoneme /y/.

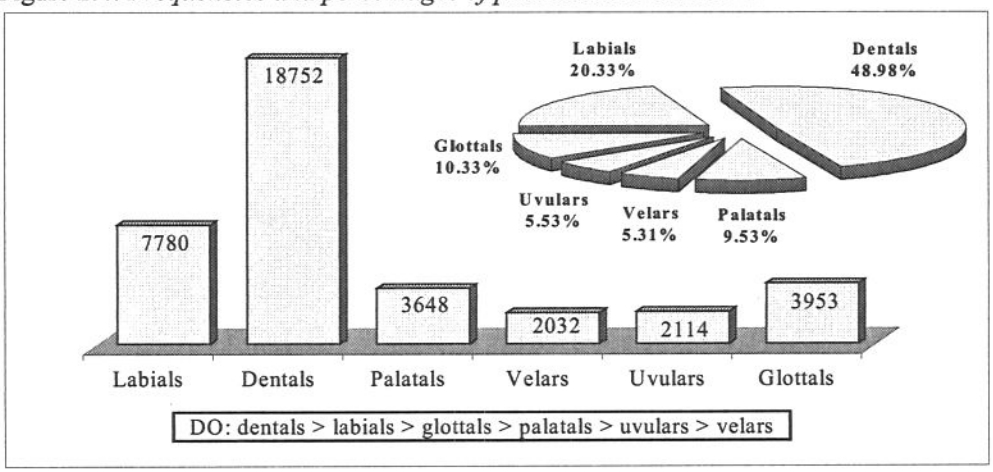
Table 2.4 presents average frequencies per consonant in each of manner subclass. We see that the two biggest subclasses, nasals and liquids, do not differ very much in their average frequencies, while the average frequency in the third group, stops, is about half of the average of nasals. Nasals have the highest average frequency and affricates the lowest.

Table 2.4: *Average frequencies of manner subclasses in the DW*

Stops	Affricates	Fricatives	Liquids	Nasals	Glide
1723.5	560.5	1204.5	3043	3282	1084

Figure 2.4 presents the frequencies and percentages for each place class. We see that dentals predominate, covering nearly half (48.98%) of all consonant occurrences in the DW. The second group, labials, takes about 20%. Of the rest, glottals and palatals take roughly 10% each, and uvulars and velars about 5% each.

Figure 2.4: *Frequencies and percentages of place classes in the DW*



But if we eliminate the effect of different group sizes and calculate average frequencies per consonant in each group, we get the results presented in Table 2.5. We see that dentals remain at the top of the list, but glottals precede labials and palatals have the lowest average frequency.

Table 2.5: *Average frequencies of place classes in the DW*

Labials	Dentals	Palatals	Velars	Uvulars	Glottals
1556	2678.8	729.6	1016	1057	1976.5

2.3. Vowel frequencies

Since a vowel, and only one vowel, can be the nucleus of a Persian syllable, the number of vowel tokens in the DW is the same as the number of syllables, which is 24135. Figure 2.5 presents vowel frequencies in the DW, together with percentages. The figure shows that the most frequent vowel, /ä/, covers nearly a third (31.05%) of all vowel occurrences, while the least frequent, /u/, occupies only 5.06% of vowel tokens.

Figure 2.5: *Frequencies and percentages of the vowels in the DW*

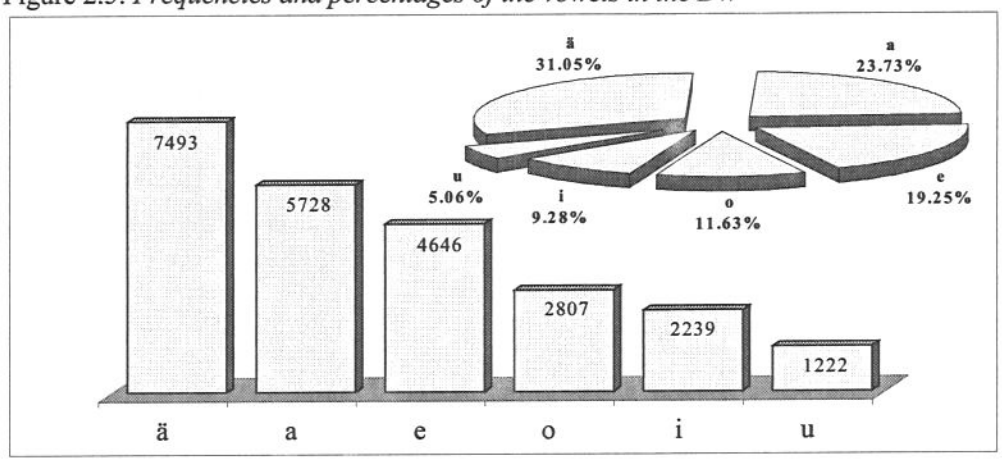
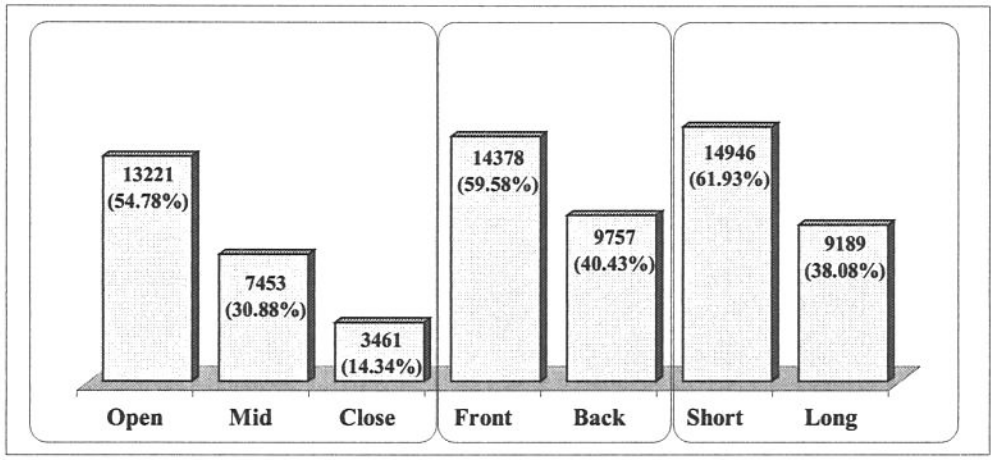


Figure 2.5 shows that the open vowels (/ä/, /a/) are the most frequent ones, the close vowels (/i/, /u/) are the least frequent, with mid vowels in between in terms of frequency. Thus, there is a clear positive correlation between the frequency and the sonority (openness) of the vowel, i.e. as the sonority increases, the frequency also increases. We also see that on each openness level, the front vowel is more frequent than the back vowel (i.e. /ä > a/, /e > o/, /i > u/).

We see these features more clearly in Figure 2.6, which presents frequencies and percentages of various vowel classes. The figure shows that open vowels cover 55% of all vowel tokens, whereas the proportion of close vowels is only 14%. Front vowels cover 60% of the tokens, back vowels 40%. Finally, short vowels (/ä,e,o/) are more frequent in the DW: their proportion is 62%, while long vowels have a proportion of 38%. Since the number of vowels is the same in each pair of groups that were compared, we can say that short vowels, front vowels, and more sonorous vowels are favoured in the DW.

Figure 2.6: *Frequencies and percentages of vowel classes in the DW*



2.4. Phoneme frequencies and etymology

As was mentioned before, the two major sources of the DW are Arabic and Persian. In addition, there are words from French, Turkish, etc. In this study, I shall give a more detailed description of Arabic, Persian and French only, since the proportions of all the others are very small.

2.4.1. Consonants

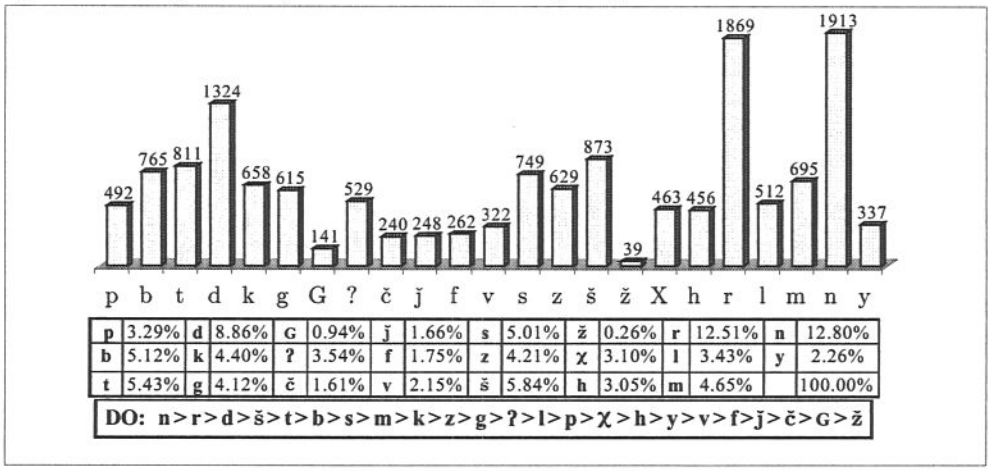
Measured by the number of consonants, the biggest contributor to the DW is Arabic and the next is Persian. Consonants from words of Arabic origin take a little more than half (52.96%) of the DW consonant tokens, and consonants from words of Persian origin nearly 40%; together these two cover about 92% of the DW material. Of the remaining consonants, words of French origin contribute a little over 5%. The number of consonant tokens according to their etymological source, plus the percentages are as follows:

	No. of C tokens	%
Arabic	20273	52.96
Persian	14942	39.03
French	1919	5.01
Others (11 languages)	1145	3.00
Total	38279	100.00

Persian

Figure 2.7 presents consonant frequencies and their percentages in the etymologically Persian words. The two most frequent consonants are the sonorants /n/ and /r/; together they cover about a quarter of all consonant tokens in words of Persian origin. These two are among the top three in the total DW as well, while /m/, which is the second most frequent in the total DW, has a relatively low frequency in etymologically Persian words. The third most frequent consonant is /d/, the voiced dental stop. It is interesting to note that in both labials and dentals, the voiced member of the pair is clearly more frequent than the voiceless one in words of Persian origin. In the data of the DW, this was true only of the labials.

Figure 2.7: *Frequencies of consonants in etymologically Persian words*



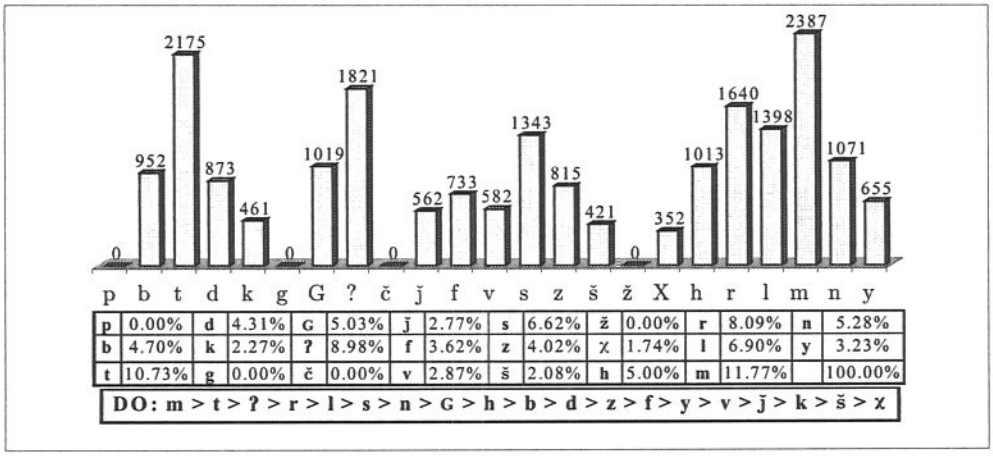
p	3.29%	d	8.86%	G	0.94%	j	1.66%	s	5.01%	ž	0.26%	r	12.51%	n	12.80%
b	5.12%	k	4.40%	ʔ	3.54%	f	1.75%	z	4.21%	χ	3.10%	l	3.43%	y	2.26%
t	5.43%	g	4.12%	č	1.61%	v	2.15%	š	5.84%	h	3.05%	m	4.65%		100.00%

DO: n > r > d > š > t > b > s > m > k > z > g > ʔ > l > p > χ > h > y > v > f > j > č > G > ž

Arabic

Figure 2.8 presents consonant frequencies and percentages in words of Arabic origin. First, we can note that four consonants have a zero frequency, i.e. they do not occur in words of Arabic origin: /p/, /g/, /č/ and /ž/. The three most frequent consonants are a nasal, /m/ and two stops, /t/ and /ʔ/. The two most frequent stops, /t/ and /ʔ/, are the same as in the DW.

Figure 2.8: *Frequencies and percentages of consonants in etymologically Arabic words*



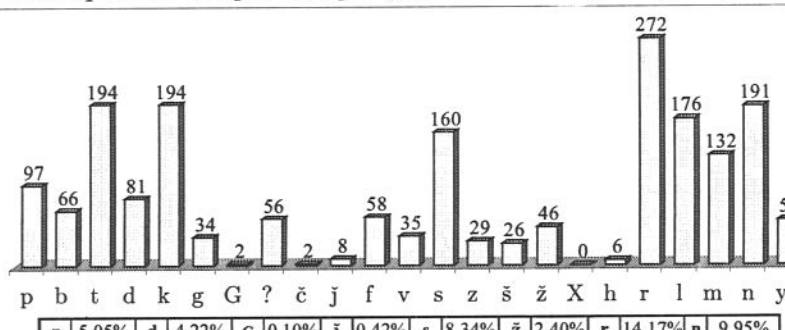
p	0.00%	d	4.31%	G	5.03%	j	2.77%	s	6.62%	ž	0.00%	r	8.09%	n	5.28%
b	4.70%	k	2.27%	ʔ	8.98%	f	3.62%	z	4.02%	χ	1.74%	l	6.90%	y	3.23%
t	10.73%	g	0.00%	č	0.00%	v	2.87%	š	2.08%	h	5.00%	m	11.77%		100.00%

DO: m > t > ʔ > r > l > s > n > G > h > b > d > z > f > y > v > j > k > š > χ

French

Figure 2.9 shows consonant frequencies with percentages in words of French origin. The top three contain the sonorant /n/ and two voiceless stops, the dental /t/ and the velar /k/. In words of French origin, voiceless stops are systematically more frequent than voiced ones. Of the fricatives, /s/ is by far the most frequent. It is interesting to note that even though only about 5% of the DW consonants come from words of French origin, more than half of the occurrences of /ž/ in the DW come from etymologically French words.[38]

Figure 2.9: *Frequencies and percentages of consonants in etymologically French words*



p	5.05%	d	4.22%	G	0.10%	ǰ	0.42%	s	8.34%	ž	2.40%	r	14.17%	n	9.95%					
b	3.44%	k	10.11%	?	2.92%	f	3.02%	z	1.51%	χ	0.00%	l	9.17%	y	2.81%					
t	10.11%	g	1.77%	č	0.10%	v	1.82%	š	1.35%	h	0.31%	m	6.88%		100.00%					

DO: r>t,k>n>l>s>m>p>d>b>f>?>y>ž>v>g>z>š>ǰ>h>G,č

[38] Figure 2.9 shows that the phoneme /h/ has a frequency of six in etymologically French words. Words in which these six cases of /h/ occurred are as follows (words enclosed between parentheses are French equivalents): /ha-šur/ (hachures), /hal-ter/ (haltére), /he-ro-ʔin/ (héroïne), /hek-tar/ (hectare), /ho-tel/ (hôtel), and /hu-ra/ (hourra). As we know /h/ does not belong to the inventory of French consonant phonemes, and the fact that it exists in words listed above, may indicate that these words have entered Persian first in written form. Persian variants of the French words I cited correspond mostly to their orthographic form rather than to the way they are pronounced in French.

Comparison

Figure 2.10 presents consonant frequencies in words of Persian, Arabic, French and other origins. Since about 53% of the consonants in the DW come from words of Arabic origin we can expect the frequency curve of the Arabic data to be above other curves, at least for several consonants. The Arabic contribution compared to the proportion of Persian is particularly strong in the stops /t/, /G/ and /ʔ/, in the fricatives /f/, /s/ and /h/, and in the sonorants /l/ and /m/. Persian, on the other hand, has a substantially bigger proportion compared to Arabic in the stops /p/, /g/ (that Arabic lacks), and /d/, in the fricative /s/ and in the nasal /n/. We see that frequencies in words of Persian and Arabic origins are complementary to some extent, especially in the sonorants: /m/ has a high frequency in etymologically Arabic words, but a low frequency in etymologically Persian words. For /n/, the situation is the opposite. In words of Persian origin, /l/ has a low frequency, but in words of Arabic origin it has a moderate frequency. A similar situation can be seen in some stops, too, e.g. /t/, /G/, /ʔ/.

Figure 2.10: *Comparison of consonant frequencies in the DW between major etymological sources (Arabic, Persian, French, and other languages)*

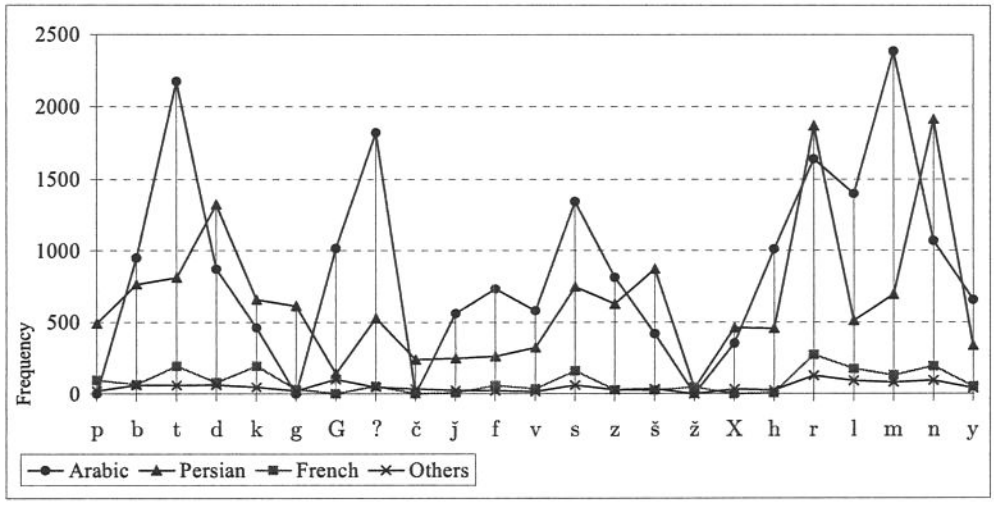
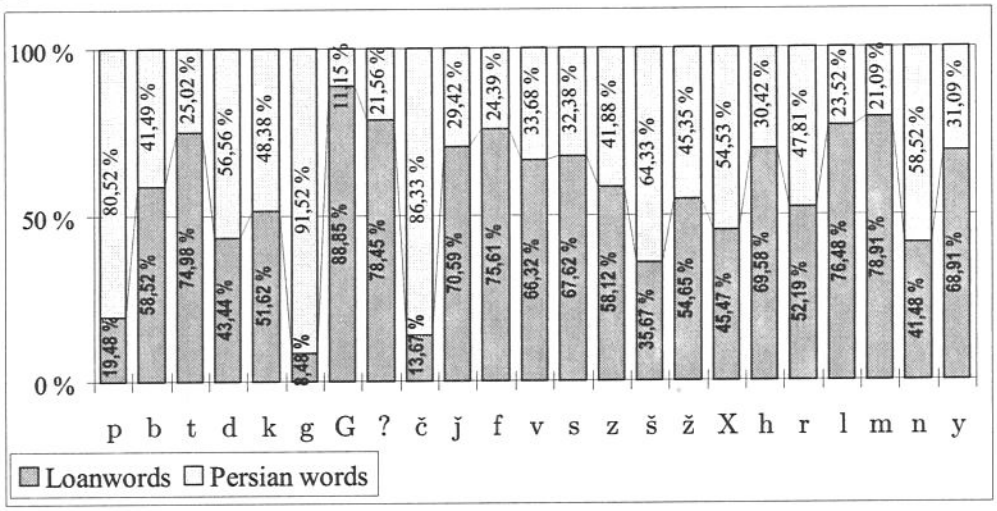


Figure 2.11 shows how much a percentage of each consonant's occurrences in the DW come from words of Persian origin and how much a percentage comes from loanwords.

Figure 2.11: *Percentages of Persian vs. foreign sources in consonant frequencies in the DW*



If the relation of the Persian vs. foreign proportion were the same for every consonant, we would expect 39% of the occurrences of each consonant to come from words of Persian origin, and the remaining 61% to come from words of foreign origin. As Figure 2.9 shows, the proportions can deviate greatly from the expected (average) values. Only /b/ and /z/ come close to those values. We see that over 80% of the occurrences of /p/, /g/ and /č/ from words of Persian origin. These are all consonants that Arabic, the biggest foreign source, lacks. On the other hand, the foreign proportion, more particularly Arabic, is around 75% or more in /t/, /G/, /ʔ/, /f/, /l/, and /m/.

Let us next look at the relationship of the native vs. foreign proportion in different consonant groups. Figure 2.12 gives the percentages in different manner subclasses. If the proportions were the same for each group, we would expect 39% of the tokens to come from Persian and 61% from foreign sources. We see that four of the groups roughly correspond to this expectation, namely stops, fricatives, liquids and nasals. The foreign proportion of the glide is more than the

average, at 61%. In affricates, the proportion of foreign sources is less than the average. This is partly due to the fact that the greatest contributor of loanwords, Arabic, lacks one of the affricates, i.e. /č/.

Figure 2.12: *Proportions of Persian vs. foreign sources in different manner subclasses in the DW*

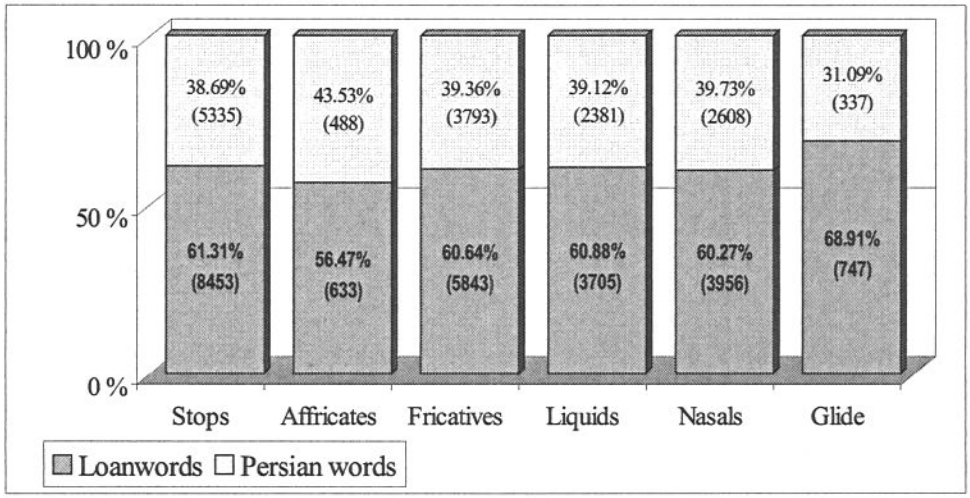
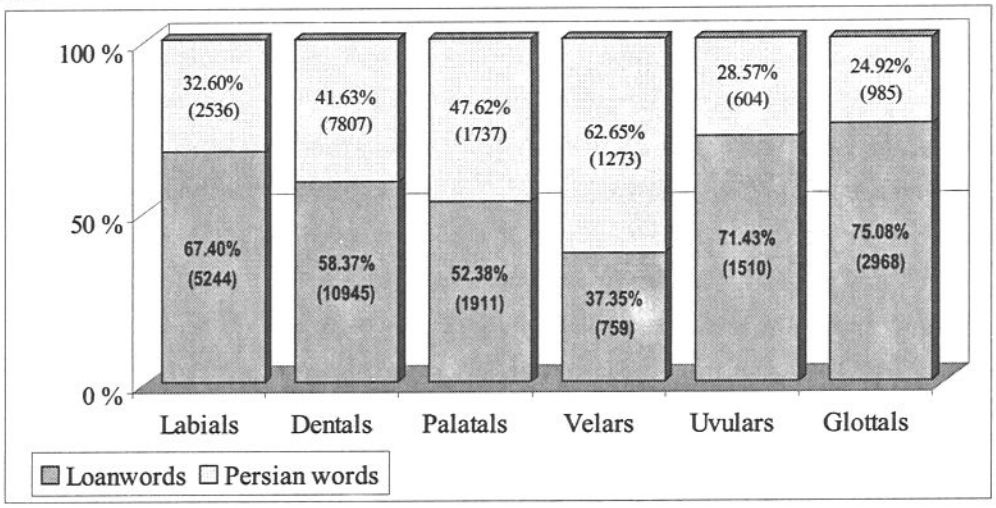


Figure 2.13 presents the percentages of Persian vs. foreign sources in different place classes. Here again 39% would come from Persian and 61% from foreign sources if the proportions were evenly distributed over the place classes. But we see that there is more variation in the proportions in place classes than there are in manner subclasses. Only the most frequent place subclass, dentals, comes close to the average values. Velars, uvulars, and glottals show greatest deviations. Over 60% of velar tokens come from etymologically Persian words, while less than 30% of uvular and glottal tokens come from words of Persian origin.

Figure 2.13: *Proportions of Persian vs. foreign sources in different place classes in the DW*



2.4.2. Vowels

As was mentioned before, the DW contains 24135 vowel tokens. They come from different etymological sources as follows:

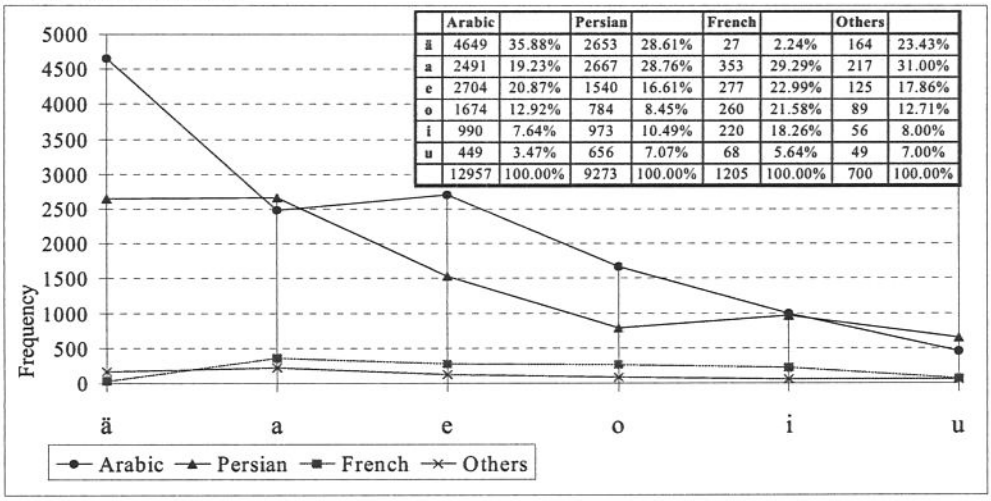
	Vowel frequency	%
Arabic	12957	53.27
Persian	9273	38.42
French	1205	4.99
Others (11 languages)	700	3.32
Total	24135	100.00

In vowel tokens, as in consonants, the highest percentage comes from words of Arabic origin. The proportion of vowel occurrences from words of Persian origin is about 38% and from foreign sources 62%.

Figure 2.14 presents the frequencies of the vowels according to the etymological group, plus the percentages. One conspicuous feature is the high

frequency of /ä/ in words of Arabic origin, it covers more than a third of vowel tokens in this group. We also see that all etymological groups tend to have higher frequencies for more sonorous (open) vowels and lower frequencies for less sonorous (close) vowels. But there are some deviations: in words of Arabic origin, the back vowel has a lower frequency than the front vowel on the same level of openness. Thus, the frequency of /a/ is very low compared to that of /ä/, and even lower than the frequency of the front mid vowel /e/. Words of Persian origin show a similar tendency for mid and close vowels. Thus, the mid back vowel /o/ has a lower frequency than the high front vowel /i/. Moreover, words of French origin deviate from the correlation between frequency and openness in the case of /ä/, which has the lowest frequency in the French group.

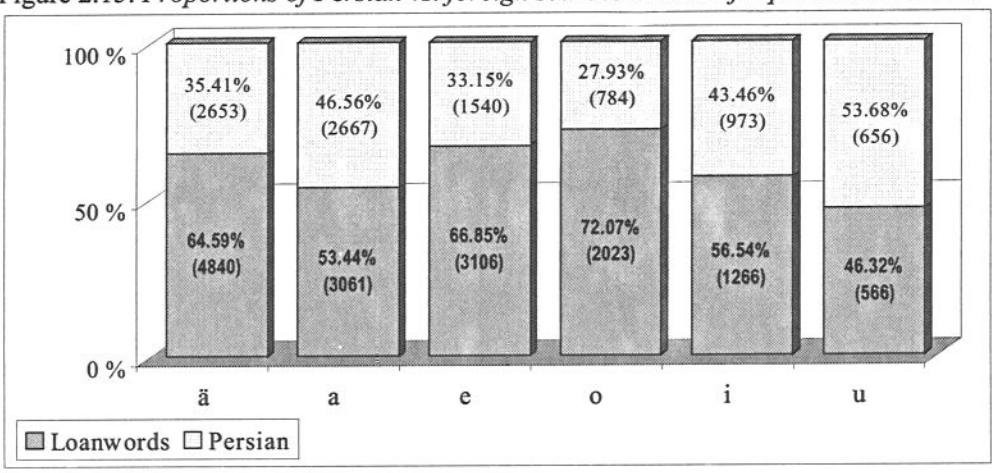
Figure 2.14: *Comparison of frequencies of the vowels between etymologically Persian words and loanwords*



	Arabic		Persian		French		Others	
ä	4649	35.88%	2653	28.61%	27	2.24%	164	23.43%
a	2491	19.23%	2667	28.76%	353	29.29%	217	31.00%
e	2704	20.87%	1540	16.61%	277	22.99%	125	17.86%
o	1674	12.92%	784	8.45%	260	21.58%	89	12.71%
i	990	7.64%	973	10.49%	220	18.26%	56	8.00%
u	449	3.47%	656	7.07%	68	5.64%	49	7.00%
	12957	100.00%	9273	100.00%	1205	100.00%	700	100.00%

We can now see what proportion of vowel frequencies comes from words of Persian vs. foreign origin. This is given in Figure 2.15. Since 62% of all vowels come from foreign sources and 38% from words of Persian origin, these figures would hold for each vowel as well, in the case that the foreign proportion was evenly distributed. But as the previous figure (i.e. 2.14) suggested, this is not the case. Only /ä/, comes close to this figure. In /a/, /i/, and /u/, i.e. in the long vowels, the proportion of etymologically Persian words is bigger, and in the mid vowels /e/ and /o/, the Persian proportion is smaller than the average 38%.

Figure 2.15: *Proportions of Persian vs. foreign sources in vowel frequencies in the DW*



Finally, let us look at Figure 2.16, which illustrates the native vs. foreign proportion in different vowel groups. If we again take 38% as the expected proportion of Persian, we see that unrounded vowels fit this figure, and that open vowels come close to it. But in mid vowels the Persian proportion is clearly smaller than expected, and in close and long vowels, the Persian proportion is bigger than expected.

Figure 2.16: *Proportions of Persian vs. foreign sources in different vowel classes in the DW*