Selected consonantal characteristics of some Finno-Ugric languages from a phonostatistical point of view[1]

Finno-Ugric linguistics has a long and fruitful development. It achieved good results in comparing Finno-Ugric languages from different aspects: the phonetical, lexical and grammatical. The methods that were applied were mostly qualitative. It was quite natural to begin with qualitative methods since linguistic units and phenomena are typically qualitative. Their only quantitative aspect consists in the frequencies with which units exemplifying their values occur. Since such units occur with a fair amount of unpredictability, they must be studied by statistical-probabilistic methods since their occurrence is of statistical-probabilistic nature (Itkonen Esa, 1980: 334 — 366). The occurrence of linguistic units in a language has a high degree of orderliness and the frequency distribution of language elements tends to preserve some shape which may be peculiar to some particular language (Zipf, 1936: III).

The application of statistical methods to Finno-Ugric studies may be considered as the second stage in Finno-Ugric studies. Numerical methods show not only the link of one language to the others inside the family, but these methods show exactly how close it is to every language of the family. It should be pointed out that though qualitative methods in Finno-Ugric studies prevailed, there were some works which used lexico-statistical methods. First we should mention the work of A. Raun where nine Finno-Ugric

[1] The author wishes to express his appreciation to Professor Nickolai G. Zagoruiko and Dr. Vicktor M. Veličko for many hours of fruitful discussions and their comments on an earlier draft of this article.

languages were studied (Raun Alo, 1956) Second, we must point out the work of E. A. Helimskij (Helimskij, 1982: 39) where the Samoyed languages were studied statistically. We are not going to discuss here the results of the above-mentioned lexico-statistical studies of Finno-Ugric and Samoyed languages, since this is done elsewhere. In short, the results are very convincing and interesting. In this article we use statistical methods in studying the phonemic frequencies of some Finno-Ugric and Samoyed languages. The comparison of these languages is based on phonostatistical results, which seem to be quite a solid foundation for any comparison. Comparing languages is a difficult task because it is hard to establish the right criterea for comparison. A well-known linguist Roman Jakobson stressed that a linguistic typology based on arbitrary selected traits cannot yield satisfactory results (Jakobson, 1958). So the method of comparison must be carefully chosen.

The method of Jiři Krámský seems to avoid the negative drawbacks of other methods of quantitative investigation, that is why it would appear quite appropriate to use this method here. A description of his method may be found in different works (Krámský, 1959, 1965, 1974). Jiři Krámský has based his method of comparing languages on the relationship between the consonant phonemes of the phonemic inventory and their relative occurrence in coherent texts. Four types of languages are distinguished by him: 1. languages overexploiting alveolars and labials; 2. languages overexploiting alveolars; 3. languages overexploiting alveolars and palatals; 4. languages overexploiting alveolars and velars.

Before the analysis of Mansi, Komi Zyryan and some other Finno-Ugric languages by J. Krámský's method, it is necessary to explain what is meant by the terms "overexploiting" and "underexploiting".

It is quite obvious that absolute frequency data giving the figures of frequency of occurrence of a particular consonant in a certain language and of another particular consonant in another certain language are not of much value for our studies. In order to be able to compare languages we must find the relative quotient of frequency of occurrence of the consonants in the invertory of a particular language in regard to their frequency of occurrence in coherent texts in this language. The assumption is that if the language exploited all consonants equally, the relative exploitation of each group of consonants (and even of individual consonants) in texts should be equal to the percentage of occurrence of these consonant groups in the phonemic inventory of this language. For example, if in the Mansi (Vogul) language there are 3 labials, 5 alveolars, 5 palatals and 4 velars, then the labials form 18 % of the inventory, the alveolars 29 %, the palatals 29 % and the velars 24 %. It should be presumed that in the case of equal distribution the texts in Mansi should also have the same percentage of labials, alveolars, palatals and velars.

J. Krámský does not state it, but in fact he uses the formula $Q_c = P_t - P_i$ where P_t is the percentage of consonants in the text, P_i is the percentage of consonants in the phonemic inventory, and Q_c is the quotient showing the exploitation of consonant groups in a particular language. A positive value for this quotient means an over-exploitation and a negative value means an under-exploitation of the consonants in question. J. Krámský is quite correct to claim that languages mostly exploit certain groups of consonants and vowels more than others, and it is this fact that gives languages their own special characteristics in one way or another and makes them sound different even if their phonemic inventories and articulation base are similar.

In this article an attempt is made to characterize Mansi and Hungarian (representing the Ugric languages), Selkup (representing the Samoyed languages), Karelian (representing the Finnic languages) and Komi Zyryan (representing the Permian languages) from the point of view of the quotient described above.[2]

Taking into account the volumes[3] of the samples in these five

[2] Mansi — the Northern dialect of Mansi which is the base of the literary Mansi language; Hungarian — the literary Hungarian language; Selkup — Sredneobskoi dialect of the Selkup language; Karelian — Ludian dialect of the Karelian language; Komi Zyrian — the literary Komi Zyrian language.

[3] It is necessary to mention where the results of statistical studies were taken from and what was the total volume of the material involved in this statistical analysis. As a matter of fact two languages: Mansi (276 418 phonemes) (Tambovcev, 1977, 1979, 1981) and Komi Zyrian (Tambovcev, in press) (80 168 phonemes) were computed by the author while the other languages' computing data were taken from literature: Hungarian (551 828 phonemes) (Jékel, Papp, 1974); Selkup (10 000 phonemes) (Morev, 1973); Karelian (62 360 phonemes) (Barancev, 1975).

languages, the data obtained after computing the material of Mansi, Hungarian or Karelian are more reliable than that of Selkup. The evidence of the theory of statistics and mathematical linguistics shows that the reliability of a sample increases with its volume (Pjotrovskij, Bektaev, Pjotrovskaja, 1977). On the whole, the larger the sample, the more closely it tends to resemble the population from which it is taken. Too small a sample would not give reliable results (Connor, Morrell, 1957). The frequency of occurrence of phonemes fluctuates greatly in different samples of small volumes, even in one and the same language and in one and the same style (Jaglom, Jaglom, 1973) and with increasing volume of a sample the error is reduced. The relation determining the sample size with regard to the admissible error can be derived (Ludvíková, Königová, 1967). It should also be noticed that the frequency of the most common (frequent) phonemes stabilizes at samples of smaller volumes, while less frequent phonemes continue to fluctuate right up to far greater volumes (Tambovcev, 1980). It must be pointed out that if the inventory of a language has more phonemes than the inventory of another language, then the sample volume for the language should be greater than that for the latter (Tambovcev, Utev, 1981).

A well-known phonologist D. Segal quite riqhtly states in his book that to his regret there are many phonostatistical studies the results of which could not be regarded as reliable or even correct since the sample volumes considered in such studies are too small and thus unreliable (Segal, 1972: 27, 68). It is a pity, but from this viewpoint the data of J. Krámský on 23 languages given in the above-mentioned works are quite unreliable because the sample volumes in all 23 languages are too small to give an opportunity for the least frequent phonemes to show their true frequencies. The same criticism may be applied to the phonostatistical works by S. Čebanov (Čebanov, 1947), by S. de Búrca (de Búrca, 1960), and some others, even such outstanding linguists as B. Bourdon. One has to agree that D. Segal's criticism of Bourdon's book "L'expression des émotions et des tendances dans le langage" (Paris, 1892), for its unreliable samples is justified, though one cannot deny that in general this book did more good than harm, and in particular greatly influenced the development of phonostatistics. As a matter of fact B. Bourdon gave tables of relative frequency of occurrence of sounds in French, German, Italian, Spanish, Russian, English and Hungarian which later were used as materials for phonostatistical models by G. Zipf (Zipf, 1929, 1932) and by G. Herdan (Herdan, 1964). Later D. Segal showed in his phonostatistical studies that the small samples of B. Bourdon (not more than 3000 sounds) did not provide for the occurrence of all the elements even once (Segal, 1972: 121 — 122). The samples of five Finno-Ugric and Samoyed languages taken for our analysis seem to be quite reliable.

Let us consider the share of the groups of consonants in the inventories of the chosen Finno-Ugric and Samoyed languages (Table 1). It should be mentioned that in the inventory of Mansi and Selkup alveolar and palatal consonants prevail equally (29,4 %, 29,4 % and 28,3 %, 28,3 % correspondingly) while in the inventories of Hungarian, Komi Zyryan and Karelian the alveolars are dominant. If we consider the percentage of labials, alveolars, palatals and velars in the inventories separately, then we can say that Karelian has more labials, Hungarian has more alveolars, Komi Zyrian has more palatals and Selkup has more velars than each of the other languages in question.

Now we should dwell on the analysis of the shares of consonants in speech. For that purpose we shall consider Table 2, which shows that in speech alveolar consonants take the greatest share in all five languages. Considering the values of the consonantal groups classified according to the place of articulation, it becomes obvious that Mansi has more labials in speech than the other languages while Karelian has more alveolars, Komi Zyryan more palatals and Selkup more velars. Functioning in speech, the consonants have different value frequencies in these five languages. So it can be noted that Mansi and Selkup correlate in the amount of labials, Karelian and Hungarian in the amount of alveolars and palatals, while Mansi and Hungarian on the one hand, Karelian and Selkup on the other, have very close values of velars. According to the values in each of the four consonantal groups these five languages may take different place in the ordered series.

So from the point of view of labiality value in the inventories the order is the following: Karelian — Selkup — Mansi — Komi Zyryan — Hungarian; alveolarity value gives the following order: Hungarian — Komi Zyryan — Karelian — Mansi — Selkup;

palatality has the following order: Komi Zyryan — Mansi — Selkup — Karelian — Hungarian and velarity the following: Selkup — Mansi — Hungarian — Komi Zyryan — Karelian. It is clear from these ordered series that Mansi is just before or after Selkup. Let us look at the ordered series constructed on the basis of labiality, alveolarity, palatality and velarity values in speech. They are: labiality: Mansi — Selkup — Komi Zyryan — Karelian — Hungarian; alveolarity: Karelian — Hungarian — Komi Zyryan — Selkup — Mansi; palatality: Komi Zyryan — Mansi — Selkup — Hungarian — Karelian; velarity: Selkup — Karelian — Hungarian — Mansi — Komi Zyryan. It is evident from these series that in most cases Mansi immediately follows or precedes Selkup. The other language that is close to Mansi from this point of view in speech and the inventories is Komi Zyryan, taking second place, while the closest — Selkup — takes first place. As for the closeness of Mansi and Hungarian from the point of view of frequency of the consonantal groups, they stand rather far apart. The only closeness that they have is in the functioning of velars (c. f. 17,4 % and 17,2 %). Before this analysis we expected Mansi and Hungarian to be the closer in these aspects than any of the other five selected languages.

Finally it would be reasonable to compare how these Finno-Ugric and Samoyed languages use certain groups of consonants in speech and in the inventory, i. e. to compare the ratio and the difference between the share of labials, alveolars, palatals and velars in speech and in the inventory. To realize it, one must consider Table 3, the first column of which reflects the values of the ratio and the second column reflects the values of the difference which are denoted by R_c and D_c correspondingly. To demonstrate the ratio R_c and the difference D_c more clearly, one must write the following formulae: $R_c = \frac{P_t}{P_i}$, and $D_c = P_t - P_i$, where P_t is the percentage of consonants in the text, and P_i is the percentage of the same consonants in the phonemic inventory.

An analysis of Table 3 shows that all five languages overexploit alveolars and underexploit palatals. All of them, exept Karelian, overexploit labials. As far as velars are concerned, only Hungarian and Karelian overexploit them, the rest underexploit them. If the values of R_c and D_c are taken into account, then these five languages should be classified as follows: 1. Mansi — a language overexploiting labials and alveolars 2. Hungarian — a language overexploiting velars 3. Komi Zyryan — a language overexploiting alveolars 4. Karelian — a language overexploiting alveolars and velars 5. Selkup — a language overexploiting labials and alveolars. Looking through this classification we can see that Mansi and Selkup are charecterized equally. It means that according to this classification Mansi and Selkup fall into one class of languages. It can also be proved graphically: the verification of this statement may be found in Figure 1 where one can see that the pattern of the distribution of labials, alveolars, palatals and velars in their inventories and speech is very much the same in both languages.

In conclusion we may remark that the similarity of Mansi and Selkup from the point of view of the distribution of the shares of consonants classified according to the place of articulation functioning in speech and in their inventories does not seem to be a mere coincidence: there must be something more basic in it. One should also remember that they are rather close territorially. There may be some contacts or influences in the past that gave rise to this phonostatistical similarity. It should be emphasized that the studies of Finno-Ugric and Samoyed languages with the help of phonostatistical methods may shed new light on the relation between them.

Outside phonostatistics and linguistics in general there is strong evidence to support our conclusion about the closeness of Mansi and Selkup: the anthropological data of G. F. Debets show the following series ordered according to the index of the face flatness (Debets, 1961:59)

1. Estonian	20,9 ± 2,7
2. Erza(mordva)	30,0 ± 2,4
3. Mari	44,0 ± 1,9
4. Khanty	67,6 ± 1,4
5. Mansi	69,6 ± 3,1
6. Selkup	70,9 ± 2,6
7. Nenets	73,9 ± 2,7

This index shows a closer relation between Mansi and Selkup than even between Mansi and Khanty, though Khanty's index value is close enough to that of Mansi.

From the point of view of the other important anthropological index shown by G. F. Debets — the index of Mongolian features (Mongolian admixture) — these two peoples — Mansi and Selkup — are also quite close (Debets, 1961:67)

1. Estonian	1,5 ± 4,5
2. Erza(mordva)	16,7 ± 4,0
3. Mari	40,0 ± 3,2
4. Khanty	79,3 ± 2,3
5. Mansi	82,5 ± 5,1
6. Selkup	84,7 ± 4,3
7. Nenets	89,8 ± 4,5

It is quite obvious from the values of this index that Selkup and Mansi are closer to each other than to Khanty, though Khanty is rather close to Mansi, but in generel Mansi and Selkup are both further from the other Finno-Ugric members of this list than from each other.

From our point of view, the application of anthropoligical, ethnographical and historical materials for the interpretation of linguistic evidence is always of great importance, and such an application actually gives weighty support to linguistical data. In this case we can only regret that we have no commeasurable data of the same kind for the other Finno-Ugric peoples discussed in this article: Karelian, Komi Zyryan and especially Hungarian.

The closeness of Mansi and Selkup from the phonostatistical and anthropological point of view may allow us in the future to consider these two languages and these two peoples from a new approach, i.e. bearing in mind that these two languages are more similar to each other than linguists used to think. Since lexics is not very reliable, more emphasis should be put on phonetic and grammatical investigations of Mansi and Selkup to search for more facts (positive or negative) for their relatedness. Maybe they used to live together and later were separeted by the Khanty, or we may suppose that originally Mansi and Selkup were more genetically related. Then they were separated by the Khanty from Mansi and were driven closer to the Nenets, who later influenced Selkup so that now we consider Selkup to belong to the Samoyed languages.

JURI A. TAMBOVCEV

Table 1. The percentage share of consonants, classified according to the place of articulation in the phonemic inventories of some Finno-Ugric and Samoyed languages, %

Language	Labials	Alveolars	Palatals	Velars
1. Mansi	17,7	29,4	29,4	23,5
2. Hungarian	12,5	58,3	16,7	12,5
3. Komi Zyrian	17,2	41,4	31,0	10,4
4. Karelian (Ludian)	30,3	36,4	24,2	9,1
5. Selkup	19,5	28,3	28,3	23,9

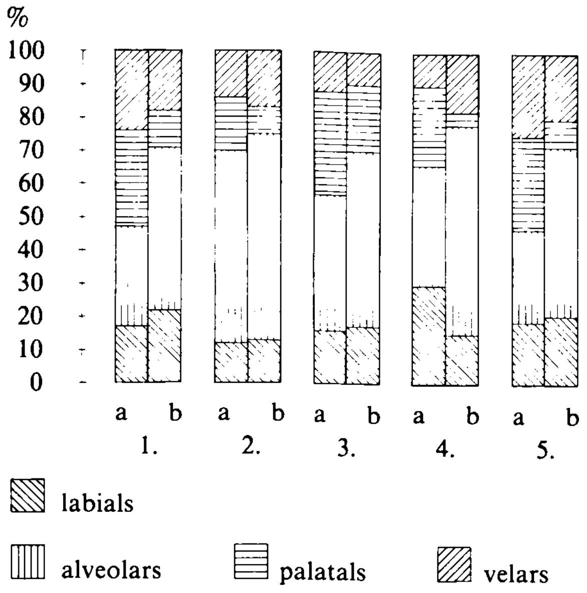
Table 2. The percentage share of consonants, classified according to the place of articulation in the speech of some Finno-Ugric and Samoyed languages, %

Language	Labials	Alveolars	Palatals	Velars
1. Mansi	22,2	49,3	11,1	17,4
2. Hungarian	12,9	62,8	7,0	17,2
3. Komi Zyrian	17,5	52,9	19,4	10,2
4. Karelian (Ludian)	15,9	63,0	2,6	18,4
5. Selkup	22,0	51,6	7,6	18,8

Table 3. The ratio and difference between the percentage shares of consonantal groups in the speech and in the inventories of some Finno-Ugric and Samoyed languages

Language	Labials		Alveolars		Palatals		Velars	
	R	D	R	D	R	D	R	D
1. Mansi	1,2	+4,5	2,5	+19,9	0,4	-18,3	0,7	- 6,1
2. Hungarian	1,0	+0,4	1,0	+ 4,5	0,4	- 9,7	1,4	+4,7
3. Komi Zyrian	1,0	+0,3	1,3	+11,5	0,6	-11,6	1,0	- 0,17
4. Karelian (Ludian)	0,5	-14,4	1,7	+26,6	0,1	-21,6	2,0	+9,5
5. Selkup	1,1	+2,5	1,8	+23,3	0,3	-20,7	0,8	- 5,1

$R = \frac{P_t}{P_i}$; $D = P_t - P_i$, where P_t is the frequency of the consonants in the speech, and P_i is the frequency of the same consonants in the inventory.



a = Inventory
b = Speech
1. Mansi
2. Hungarian
3. Komi Zyrian
4. Karelian (Ludian)
5. Selkup

Figure 1. The percentage share of consonantal groups in the phonemic inventories and the speech of some Finno-Ugric and Samoyed languages.

REFERENCES

Barancev A., 1975. Fonologičeskie sredstva l'udikovskoj reči. Leningrad, p. 280.

Bourdon B., 1892. L'expression des émotions et des tendances dans le langage. Paris.

de Búrca S., 1960. Irish phoneme frequencies. Orbis, Vol. IX, No. 2.

Čebanov S. G., 1947. O podčinenii rečevyx ukladov "indo-evropejskoj" gruppy zakonu Puassona. Doklady AN SSSR, t. LV, No. 2, pp. 103 — 106.

Debets G. F., O put'ax zaselenija severnoj polosy russkoj ravniny i vostočnoj pribaltiki. Sovetskaja Ėtnografija, No. 6, 1961, pp. 51 — 69.

Helimskij E. A., 1982. Drevnejšie vengersko-samodijskie jazykovye paralleli (Lingvističeskaja i ėtnogenetičeskaja interpretacija). Nauka, Moskva.

Herdan G., Quantitative Linguistics. London 1964.

Itkonen Esa, 1980. Qualitative vs. Quantitative Analysis in Linguistics. Evidence and argumentation in Linguistics. Edited by Th. A. Perry. Berlin, Walter de Gruyter, pp. 334 — 366.

Jakobson R., 1958. Typological studies and their contribution to historical comparative linguistics. Proceedings of the 8th International Congress of Linguistics. Oslo, p. 17.

Jaglom A. M., Jaglom I. M., Verojatnost' i informacija. Moskva 1973, p. 58.

Jékel P., Papp F., 1974. Ady Endre összes költői műveinek fonémastatisztikája. Budapest.

Krámský J., 1959. A quantitative Typology of Languages. Language and Speech. Vol. 2, part 2, pp. 72 — 85.

— 1965. Some statistical Observations on the Role of the Place of Articulation in Languages. Philologica Pragensia 8 (47), 2 — 3, pp. 245 — 250.

— 1974. Notes on Quantitative Typology of Languages. Acta Universitatis Carolinae — Philologica 5, Linguistica Generalia 1. Praha, pp. 147 — 149.

Ludvíková M., Königová M., 1967. Quantitative Research of Graphemes and Phonemes in Czech. The Prague Bulletin of Mathematical Linguistics, No. 7, pp. 15 — 29.

Morev J. A., 1974. Zvukovoj stroj sredneobskogo (laskinskogo) govora sel'kupskogo jazyka. Tomsk.

Raun Alo, 1956. Über die sogenannte lexikostatistische Methode oder Glottochronologie und ihre Anwendung auf das Finnisch-Ugrische and Türkische. Ural-Altaische Jahrbücher XXVIII.

Segal D. M., 1972. Osnovy Fonologičeskoj Statistiki. Nauka, Moskva.

Tambovcev J. A., 1977. Nekotorye xarakteristiki raspredelenija fonem mansijskogo jazyka. SFU XIII, pp. 195 — 198.

— 1979. Raspredelenie glasnyx fonem v mansijskoj poesii. SFU XV, pp. 164 — 167.

— 1980. Častotnye xarakteristiki glasnyx pervogo sloga mansijskogo jazyka. Zvukovoj stroj Sibirskix jazykov. Novosibirsk, pp. 72 — 73.

— 1981. Zakonomernosti častotnogo funkcionirovanija dolgix i kratkix glasnyx v udarnyx i neudarnyx slogax mansijskogo slova. SFU XVII, pp. 105 — 109.

— 1981. Szótagtipusok az északi vogul nyelvjárásban. Nyelvtudományi Közlemények 83, 1, pp. 133 — 138.

Tambovcev J. A., Utev S. A., 1981. Zavisimost' veličiny častot mansijskix glasnyx 1-go sloga ot veličiny obyoma vyborki. Teoretičeskie voprosy fonetiki i grammatiki jazykov narodov SSSR. Novosibirsk, pp. 104 — 105.

Zipf G. K., 1929. Relative Frequency as a Determinant of Phonetic Change. Harvard Studies in Classical Philology, No. 40. Cambridge, Mass.

— 1932. Selected Studies of the Principle of Relative Frequency in Language. Cambridge, Mass.

— 1936. The Psycho-Biology of Language. An Introduction to Dynamic Philology. London.