PHONOSTATISTICAL CHARACTERISTICS OF THE ESTONIAN LANGUAGE

JUHAN TULDAVA

Dept. of Applied Linguistics
Tartu State University
Tartu, Estonia, USSR 202400

ABSTRACT

The paradigmatic as well as the syntagmatic (positional) relations between the phonemic units of the Estonian language are examined from the quantitative point of view. The results of the investigation are compared with analogical data from some other languages (particularly Finnish and Hungarian).

THE INVENTORY

The phoneme inventory of the Estonian language contains 9 vowels /a e i o u õ ä ö ü/ and 17 consonants /p t t́ k f h j l ĺ m n ń r s ś š v/ [1; 2]. All these phonemes may be short or long. The long monophthongs and long consonants are considered to be single phonemes. There are 36 diphthongs in Estonian [3] but phonologically they are treated as sequences of two vowels. All nine Estonian vowels contrast in stressed position but in unstressed position only four of them (/a e i u/) occur in the normal system (the literary language). The first component of an Estonian diphthong may be any of the nine vowels but the second component has to be chosen out of the first five vowels /a e i o u/, not all of these combinations being acceptable [3].

In orthography the long vowels are marked with two graphemes representing the same quality (e.g. maa /mā/ 'land, country'). The long consonants may be marked with two graphemes (e.g. linn /lin̄/ 'town') or sometimes with one grapheme (linlane /lin̄lane/ 'town-dweller').

All stops in Estonian are unvoiced, the distinction is made between short and long stops (lenis and fortis on the phonetical level). The short stops may be marked with the graphemes b, d, g or p, t, k (e.g. viga /vika/ 'mistake' and kord /kort/ 'order') The long stops are usually marked with two graphemes (pp, tt, kk) or in some positions with only one grapheme (pikk /pik̄/ 'long' and piklik /pik̄lik̄/ 'oblong'). For more detailed analysis the quantity alternation of the Estonian language has to be considered (short, long, overlong).

The phonology of a language cannot be regarded as complete if it does not take into account some basic quantitative (statistical) features of the system and the functioning of its units in speech (text). For instance, the number of vowels in a phonemic system indicates the degree of "vocalism" (Vokalhaltigkeit) and may be regarded as a typological characteristic of a language [4; 5]. But even more important for the phonostatistical study of languages is the investigation of the frequency of occurrence of phonemic units in text.

TEXT FREQUENCIES

Our study is based on a corpus of texts of the contemporary Estonian language (55 % of fiction and 45 % of non-fiction) with a total of about 150,000 running phonemes. The results of the statistical investigation will be given in a simplified form: the frequencies of short and long phonemes (e.g. /a/ and /ā/) are counted summarily and so are the frequencies of the non-palatalized and palatalized forms of the consonants /t l n s/. In this case the total number of phonemes is 22.

If we group these phonemes according to their occurrence we can distinguish three main groups constituted by phonemes of relatively high frequency ($p \geqslant 6$ %), medium ($6 < p < 2$) and low frequency ($p < 2$):

a	12.2	n	4.6	j	1.9
t	11.9	m	4.0	h	1.7
e	11.0	o	3.1	ä	1.3
i	9.5	r	2.9	õ	1.3
s	9.0	p	2.6	ü	0.9
k	7.3	v	2.3	ö	0.2
l	6.2			f	0.05
u	6.0			š	0.05

In full accord with other linguistic levels the functioning of the phonemic system in text reveals the tendency of concentration and dispersion of its units: we can distinguish the "core" (nuclear part) of the system, the intermediate part, and the "periphery". The three most frequent phonemes /a t e/ cover 35.1 % of the Estonian text, the eight most frequent ones - 73.1 %, and the ten most frequent phonemes - 81.7 %.

The phenomenon of concentration and dispersion is well-known in lexical statistics where the statistical distribution of the units may be expressed analytically by the so-called Zipf's law in the form of a power function. Unlike the lexical level with a very large number of units the phonemic level with its limited inventory is not submitted to Zipf's law but to logarithmic or exponential law of growth (or decrease). This can be demonstrated on our experimental material (Fig. 1): there is evidently a linear relation between the logarithm of probability (relative frequency) of a phoneme and its place (rank) in the hierarchy of units. In other words, it means exponential dependence

$$p_i = ae^{-bi} \qquad (1)$$

where p_i is relative frequency, i - rank, a and b^1- constants, and e - the base of natural logarithms. In our example a ≈ 17 and b ≈ 0.15.

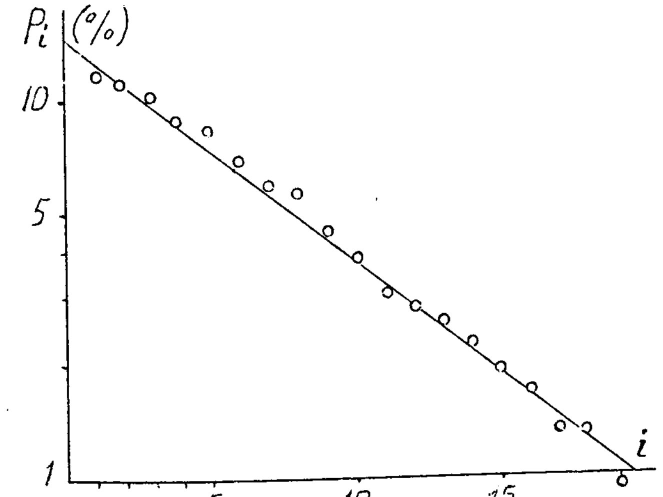

Fig. 1. Linear relation of rank (i) and the logarithm of occurrence probability of a phoneme (ln p_i).

The concrete form and the values of the constants in the formula approximating the empirical curve may serve as typological characteristics of a language. It may be added that the principle of concentration and dispersion of units in any concrete manifestation is considered to express a universal law which is peculiar to certain self-regulating systems in social life.

Another method of estimating the state of the functioning system as a whole is the measurement of the entropy of the system. The entropy of phoneme frequencies is defined as

$$H = - \sum_{i=1}^{k} p_i \log_2 p_i \qquad (2)$$

where H marks entropy, p_i is the probability (in the empirical case - the relative frequency) of the phoneme in a system of k phonemes; $\log_2$ means logarithm with base 2.

For the simplified Estonian system with 22 phonemes we get H = 3.9063. In terms of the theory of information, we can say that the entropy per phoneme occurrence is 3.9063 bits of information. Actually the entropy measures the degree of "equidistribution" of the phonemes in text. For comparison with other results we have to compute the relative entropy

$$H_{rel} = \frac{H}{H_o}$$

where $H_o = \log_2 k$. It is necessary in cases where the compared systems have different numbers of elements. For instance, we can compare our results with the results of other investigations [6] (Table 1).

Table 1
Entropy of phonemic systems

Language	k	H	H_o	H_{rel}
Estonian	22	3.9063	4.4594	0.8760
Hungarian	39	4.6028	5.2854	0.8709
German	33	4.4435	5.0444	0.8809
English	39	4.7098	5.2854	0.8911
Russian	41	4.8257	5.3576	0.9007

The smaller the value of H_{rel}, the more compressed is the series of phonemes against that of equidistribution. In this respect Estonian and Hungarian, having relatively low values of H_{rel}, differ discernibly from other languages under examination.

However, if we compare the statistical distribution of the frequencies of concrete phonemes e.g. in Estonian and Hungarian, we may find both resemblances and essential differences. In Hungarian the eight most frequent phonemes in texts of fiction are /e a t n l k m r/ [7]. Five of them coincide in Estonian and Hungarian and the three most frequent ones are the same (/a e t/). But there are differences in the distribution of medium and low frequency phonemes and in phoneme systems on the whole.

As to Finnish it must be noted that there are some pecularities in the distribution of phonemes in Finnish texts that make the difference between the two close cognates - Finnish and Estonian - remarkable enough. The most frequent phonemes in Finnish texts are /a n i e t s ä k o l/ [8]. The most striking difference lies in the frequency of occurrence of the phoneme /n/. In Finnish it occupies the second place with the relative frequency of about 10 %

(in Estonian /n/ is on the 9th place with the frequency of 4.6 %). This is to a great extent due to the high frequency of occurrence of final /-n/ in common words where Estonian has lost the final consonant in the course of historical development, e.g. Finnish niin - Estonian nii 'so', kuin - kui 'when', paljon - palju 'much', or in genitive and illative forms, e.g. Finnish jalan, jalkaan - Estonian jala, jalga 'foot'. Here we can see real interdependence which exists among separate language levels where a quantitative change in the phonological system is parallel to or motivated by the structural needs and demands of some higher level of the same language, viz. of its morphological level (cf. [9]).

PHONEME CLASSES

At the first stage of classification the phonemes are divided into two large classes: vowels (V) and consonants (C). In phonostatistical works the ratio C:V is considered to be an important typological characteristic of languages [10]. In Estonian texts the ratio is 54.5:45.5 (%), or 1.20, i.e. the consonants exceed the vowels by 20 %. This value (1.20) can be compared with the corresponding values of the ratio in other languages:

	C		V		
Finnish	52.3	:	47.7	=	1.10
Italian	54	:	46	=	1.17
Lithuanian	56.3	:	43.7	=	1.29
Ukrainian	57.8	:	42.2	=	1.37
Russian	58	:	42	=	1.38
Hungarian	58.6	:	41.4	=	1.41
Czech	58.7	:	41.3	=	1.42
Polish	58.8	:	41.2	=	1.43
German	59.7	:	40.3	=	1.49
English	60	:	40	=	1.52

As to the vowel phonemes we can further divide them into several classes according to their phonetic properties. The frequencies of occurrence of vowel classes in Estonian texts are given in Table 2 (unr - unrounded, ro - rounded).

Table 2
The vowel system: frequencies in text

	Front		Back		Total
	unr	ro	unr	ro	(%)
High	i 20.8	ü 2.0	õ 2.9	u 13.2	38.9
Mid	e 24.2	ö 0.4	-	o 6.8	31.4
Low	ä 2.9	-	a 26.8	-	29.7
Total (%)	47.9	2.4	29.7	20.0	100.0
	50.3		49.7		

As can be seen, the front and back vowels in Estonian texts are equally distributed (about 50:50 %). In Finnish and Hungarian the front:back ratio is 52:48, the same in Italian, but for instance in Slovak it is 43:57, and in Sanskrit texts 20:80.

The relation of the frequency of short vowels to the frequency of long vowels in Estonian texts is 92 to 8 %. The same relation characterizes Finnish texts, whereas in Hungarian the long vowels occur more often and the ratio "short:long" is 80:20 (%).

The classification of consonants according to the manner of articulation and according to the place of articulation are brought together in the synoptic Table 3.

Table 3
The consonant system: frequencies in text

	Labial	Alveodent		Palatal	Velar	Total (%)
		non-pal	pal			
Stops	p 4.8	t 21.8	t′	-	k 13.4	40.0
Fricat	f 0.1	s 16.5	s′	š 0.1	h 3.1	19.8
Nasals	m 7.3	n 8.5	n′	-	-	15.8
Laterals	-	l 11.4	l′	-	-	11.4
Trills	-	r 5.3	-	-	-	5.3
Semi-vowels	v 4.2	-	-	j 3.5	-	7.7
Total (%)	16.4	63.5		3.6	16.5	100.0

Two parallel sets of alveodentals (except /r/) can be distinguished: non-palatalized and palatalized consonants. It has been ascertained that except in case of automatic palatalization before /i/ and /j/ the palatalized consonants /t′ s′ n′ l′/ cover only 0.15 % of all running phonemes in Estonian texts [11].

The identification of long consonant phonemes in a running text is problematic in some cases. We estimate that about 17 % of consonants are long and 83 % short.

As a whole, the quantitative distribution of phonemes in Estonian texts can be illustrated in the following manner:

$$\begin{array}{lrl} & \text{Vowels } 45.5 & \left.\right\} \text{"Resonants"} \\ \text{"Consonants"} & \left\{\text{Sonorants } 21.9\right. & \left.\right\} \ (67.4) \\ \quad (45.5) & \left\{\text{Obstruents } 32.6\right. & \\ & \overline{\quad 100.0\ (\%)} & \end{array}$$

POSITION ANALYSIS

The phonemes occur with different frequencies in different positions of the word. In principle, initial, medial and final positions can be distinguished.

In the Orthological Dictionary of the Estonian language [12] in 115,000 entries the most frequent initial phonemes are (%): /k/ 17.4, /p/ 11.8, /t/ 10.3, /s/ 9.2, /v/ 6.3, /m/ 6.3, /l/ 6.1, /a/ 5.5, /r/ 5.4, /h/ 4.2. Among the ten most frequent phonemes there is only one vowel (/a/). On the whole, the vowels make up 15.5 and the consonants 84.5 per cent of all initial phonemes in the dictionary.

On the text level the most frequent initial phonemes are (%): /k/ 14.1, /t/ 9.9, /s/ 8.8, /m/ 8.3, /p/ 7.6, /o/ 7.0, /v/ 6.5, /e/ 6.4, /j/ 5.6, /a/ 5.3 followed by /n, l, h, r, i, ü, õ, u, ä, ö/ and the "foreign" phonemes /f/ and /š/. The five most frequent initial phonemes are all consonants and they cover about 50 % of all word initial phonemes in the text.

The over-all distribution of phoneme classes in initial positions is presented in Table 4.

Table 4
Distribution of initial phonemes

Phoneme class	Dictionary		Text	
Obstruents: stops	39.5	55.1	31.6	43.4
fricat.	15.6		11.8	
Sonorants: nasals	9.4	29.4	12.9	31.4
laterals	6.1		3.8	
trills	5.4		2.6	
semivowels	8.5		12.1	
Vowels	15.5		25.2	
Total (%)	100.0		100.0	

In Finnish the vowels cover 20 % and the consonants 80 % of all word initial positions in the text. The most frequent initial phonemes are /j s k h t m o v p e/. Compared with Estonian the phonemes /j/ and /h/ are of exceptionally frequent occurrence in initial positions.

As the structure of the stressed syllable is somewhat different from that of the unstressed syllables, it is expedient to examine the frequency distribution of vowels in the nuclei of stressed syllables separately (including the nuclei of monosyllabic words): single vowels 88.0 (76.5 % short and 11.5 % long) and diphthongs (i.e. 2-vowel sequences) 12.0 %. The frequencies of single vowels: /a/ 20.3, /e/ 19.0, /o/ 12.1, /i/ 11.2, /u/ 9.0, /ä/ 5.9, /õ/ 5.5, /ü/ 4.0, /ö/ 1.0. The most frequent diphthongs: /ei/ 2.7, /ea/ 1.7, /õi/ 1.6, /ui/ 1.3, /äi/ 0.9, /ai/ 0.7.

The distribution of word final phonemes reflects the morphological structure of the language and therefore the frequencies of final phonemes are considered to be specific for each language. In Estonian texts the most frequent final phonemes are: /a/ 21.1 % (of all final phonemes in the text), /t/ 20.5, /e/ 13.6, /s/ 13.4, /i/ 12.9. These five phonemes cover 81.5 % of all word endings in the text. They are followed by the less frequent phonemes: /l/ 4.7, /u/ 4.4, /n/ 3.1, /p/ 2.3, /k/ 1.7, /r/ 0.6, /v/ 0.4. Due to the restrictions in the distribution of vowels in unstressed syllables the phonemes /o õ ä ö ü/ are extremely rare in word endings (total 0.3 %) and so are the phonemes /h/ and /f š/ (the last two occur only in foreign or recent loan words); the three phonemes have a total frequency of 0.1 %. The distribution of phoneme classes in final position: obstruents 38.0 %, sonorants 9.7 %, and vowels 52.3 %.

In Finnish the most frequent phonemes in word final position are: /n a ä i t e s o u y/. The final /n/ covers almost 30 % of all word endings in the text.

On the basis of the frequencies of phonemes in initial and final positions their relative frequencies in medial positions can be calculated.

Some other traditional problems in phonostatistics, such as the valency fields of phonemes, phonotactic features and frequencies of phoneme sequences and syllables, word length, etc., as well as a more detailed quantitative analysis of phonological data - including stress and quantity - require special discussion.

REFERENCES

[1] T.-R. Viitso. Läänemeresoome fonoloogia küsimusi. Tallinn: KKI, 1981.

[2] A. Eek. Kvantiteet ja rõhk eesti keeles (II). Seisukohavõtt. - Keel ja Kirjandus 1987, nr. 3, 153-160.

[3] H. Piir. Acoustics of the Estonian diphthongs. - Estonian Papers in Phonetics 1982-1983. Tallinn: KKI, 1985, 5-96.

[4] G. Altmann, W. Lehfeldt. Allgemeine Sprachtypologie. Prinzipien und Meßverfahren. München: W. Fink Verlag, 1973.

[5] U. Strauß. Struktur und Leistung der Vokalsysteme./Quantitative Linguistics, vol. 4. Bochum: Brockmeyer, 1980.

[6] P. Zörnig, G. Altmann. The entropy of phoneme frequenciesand the Zipf-Mandelbrot law. - In: Glottometrika 6./Quantitative Linguistics, vol. 25. Bochum: Brockmeyer, 1984, 41-47.

[7] F. Papp. Lingvostatistika i vengerskij jazyk. - Acta et Commentationes Universitatis Tartuensis, vol. 518, Tartu, 1980, 15-37.

[8] V. Setälä. Suomen kielen dynamiikka I. Helsinki: SKS, 1972.

[9] J. Vachek. Prague phonological studies today. - Travaux linguistiques de Prague I. Prague: Academia, 1966, 7-20.

[10] Yu. Tambovcev. Konsonantnyj koefficient v jazykax raznyx semej. Leningrad, 1986.

[11] M. Hint. Häälikutest sõnadeni. Tallinn: Valgus, 1978.

[12] Õigekeelsussõnaraamat./Toimet. R.Kull, E. Raiet. Tallinn: Valgus, 1976.