Linguistica Uralica XXXIX 2003 4

LIINA LINDSTRÖM (Tartu), *KARL PAJUSALU* (Tartu—Helsinki)

CORPUS OF ESTONIAN DIALECTS AND THE ESTONIAN VOWEL SYSTEM*

Abstract. This article consists of two parts: an introduction to the principles and the current state of the corpus of Estonian dialects, and a presentation of the main characteristics of the vowel systems of Estonian dialects based on statistical analysis of the data in the dialect corpus. First, the starting points and problems that had to be taken into consideration when compiling the corpus are introduced, and the development of the project up to now reviewed. Thereafter, the state of the dialect corpus as it stands in October 2003 will be described together with the principles of tagging and a frequency study of the dialect vocabulary carried out on the basis of the corpus. The characterization of the vowel systems of Estonian dialects will be presented according to the general distribution of the distinctive features.

Introduction

The territory where Estonian is spoken is quite small but there are large differences between traditional dialects. Researches of Estonian dialects have classified at least eight main dialects and over hundred sub-dialects or parish dialects (see Pajusalu, Hennoste, Niit, Päll, Viikberg 2002; Pajusalu 2003).

The traditions of Estonian dialectology are also rather long. Andrus Saareste introduced dialect geography to Estonia already in the 1920s and compiled several Estonian dialect atlases starting from the late thirties. These were based on huge dialect data collections. Up to now, the Estonian dialect archive at the Institute of Estonian Language in Tallinn contains more than two million data units of dialect words, over 2900 hours of sound recordings with examples of each Estonian sub-dialect, and several thousand pages of transcribed texts. There are additional collections of Estonian dialect data at the universities of Tartu and Tallinn.

The amount and scope of comparative studies on Estonian dialect phonology and grammar has, however, been rather limited (an outstanding exception is Tauli 1956) because of the lack of a united data source for such kind of analysis. For facilitating such studies, the University of

* This study has been carried out as part of the projects No. 4192 and 4404 of the Estonian Science Foundation.

Tartu and the Institute of Estonian Language in Tallinn started a joint project for compiling an electronic corpus of Estonian dialects in 1998 (see also Lindström, Lonn, Mets, Pajusalu, Teras, Veismann, Velsker, Viikberg 2001). The main aim of this corpus is to enable the study the phonological and grammatical structure of Estonian dialects by means of electronic data processing. The corpus is planned to contain digitized sound recordings and electronic text versions of recordings from all Estonian dialects and main sub-dialect groups within the dialects.

1. Current state of the corpus

1.1. Classification of Estonian dialects in the corpus

We have followed the most detailed classification of Estonian dialects according to which the North Estonian dialect group includes four dialects — Insular, Western, Mid, and Eastern dialects; the South Estonian dialect group consists of three dialects — Mulgi, Tartu, and Võru, including Setu dialect; the North-Eastern Coastal Estonian group includes the Coastal and North-Eastern dialects (see figure 1). Because of the exceptional position of Setu we have treated it as the fourth main dialect of South Estonian and thus there are ten dialects distinguished in the corpus.

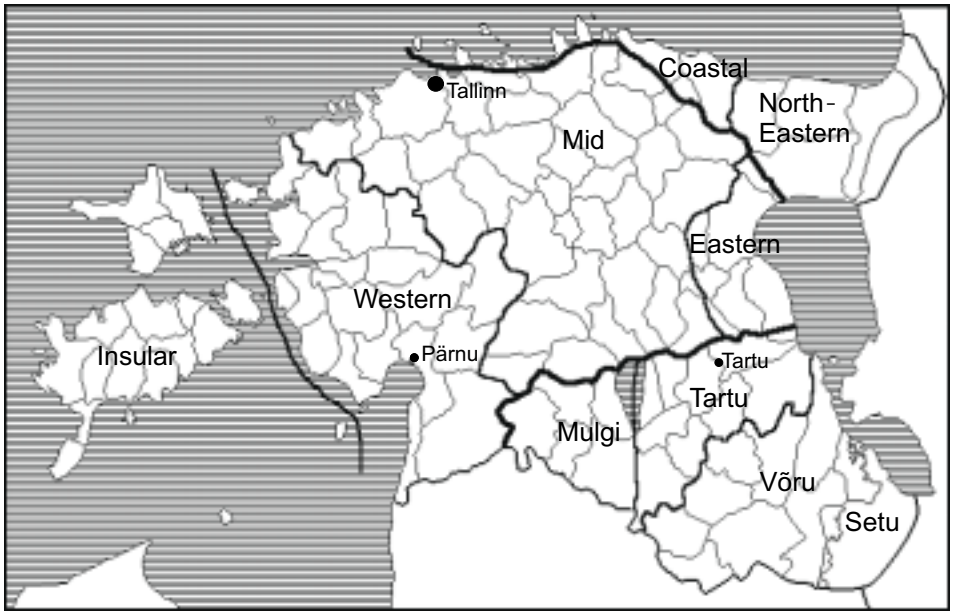

Figure 1. **Estonian dialects.**

1.2. Basic statistics of the current state of the corpus

Within the ten dialects we have determined the main sub-dialect groups and from each of these we have tried to choose examples of central parish dialects. In October 2003, there were about 456 000 text words in the corpus and the work is still in progress. Our aim is to compile a collection of at

least half a million text words of archaic dialects for the first stage of the corpus (by the end of 2003). The second stage of the corpus should include more archaic parish dialects (theoretical maximum is about 2 million words) and newer data from Estonian vernaculars.

At the moment the corpus includes archaic dialect data from all the dialects and 34 sub-dialects. Table 1 shows the number of text words for each dialect in the corpus.

Table 1

The number of words in the corpus (October 2003)

Dialect	Sub-dialects	Number of words
Coastal	Jõelähtme, Kuusalu	43 905
North-Eastern	Jõhvi, Lüganuse	36 550
Insular	Kihelkonna, Kihnu, Käina, Mustjala, Pühalepa	88 764
Western	Häädemeeste, Mihkli, Varbla	40 625
Mid	Juuru, Jüri, Keila, Pilistvere, Viru-Jaagupi, Väike-Maarja	45 179
Eastern	Kodavere, Torma	20 499
Mulgi	Halliste, Karksi, Tarvastu	23 903
Tartu	Kambja, Nõo, Otepää, Rõngu, Võnnu	67 682
Võru	Hargla, Põlva, Räpina, Urvaste, Vastseliina	46 574
Setu	Northern Setu, Western Setu	42 219
Total		455 900

Current data of the corpus is based on the oldest sound recordings of the dialects. These are interviews on various topics. The first records originate from 1938. The largest amount of text was recorded in the 1960s and 1970s (see table 2). Most of the speakers were born in the second half of the 19th century (see table 3).

Table 2

Data collection periods

Year	Number of recordings
1938	5
1957—1959	17
1960—1969	60
1970—1979	31
1980—1986	8
unknown	1
Total	122

Table 3

Years of birth of the informants

Year	Number of informants
1865—1869	7
1870—1879	40
1880—1889	44
1890—1899	20
1900—1909	13
1910—1919	7
Total	131

The corpus is in fact a text collection of spontaneous spoken language. We have taken into account special features of speech and transliterated all discourse particles, word repetitions, corrections, pause-fillers, and so on. The interviewer's text has also been transliterated.

The texts are presented in two versions. At first, there are texts transliterated in the standard Finno-Ugric phonetic transcription. The reason why we have used the Finno-Ugric transcription, which is unknown for most researchers of other language families, and not the International Phonetic Alphabet (IPA), lies in the tradition of Estonian dialectology. All the old texts of Estonian dialects based on sound recordings were transcribed in the Finno-Ugric transcription. In the future, it will be possible to modify the texts additionally into an IPA version because the current version of Finno-Ugric transcription is sufficiently precise for that.

It is already possible to carry out various types of phonetic research on this version of the corpus. For example, Karl Pajusalu, Merike Parve, and Pire Teras have studied the prosody and vowel system of Southern Estonian dialects on the basis of the corpus (Pajusalu, Parve, Teras 2001; Parve 2003; Teras 2003).

The second version of the texts is available in simplified transcription. For that the phonetic texts have been converted automatically into simple txt-format. The version is aimed to be a basis for studying grammar but marks also several features of spoken language, such as pauses and co-articulations. Compound words, interviewer's text and commentaries are marked with special symbols.

1.3. Morphosyntactic tagging in the corpus

The tagging of morphosyntactic categories has already started. The dialect corpus is a multi-lingual corpus by its nature, or more precisely, it is a corpus of languages without existing standards and complete knowledge about the linguistic structures. For this reason the tagging of parts of speech must be relatively open. It has to be possible to make corrections and introduce new categories.

We have started working out principles of morphological tagging following the example of the corpus of Finno-Ugric languages compiled at the University of Helsinki (Suihkonen 1998). In comparison with Helsinki corpus, we have added some parts of speech, e.g. discourse particles and onomatopoetic words, and we have determined several sub-classes, such as pro-adjectives and pro-adverbs.

A more appropriate example of morphological tagging for our purposes is being compiled at the Institute of Estonian Language in Tallinn where a morphological data base of Southern Estonian morphology is being put together (see http://www.eki.ee/dict/hargla/). Our aim was to make it possible to join the two databases in the future and therefore most of the tags are the same in these two databases. We have added some categories which are not used in Southern Estonian (e.g. possessive suffixes) and some categories specific to spontaneous speech (e.g. discourse particles). We have been able to distinguish between interrogative words and adverbs and relative-interrogative pronouns (where they traditionally belong) and we have separated pro-adjectives and pro-substantives.

For all the text words the following information is given (see table 4).

Table 4

Information fields

*SNE:	precise phonological shape of the word
FRA:	phrase where the word appears
*MSN:	a lemma or base form for the word (in most cases it is an entry for the word given in the Dictionary of Estonian Dialects)
TAH:	the meaning (if it is different for the standard meaning of the entry)
*SLK:	part of speech
*MRF:	morphological categories
* —	filling is obligatory

Each text file is marked by the following information:
1) dialect; 2) sub-dialect; 3) village; 4) informant's name and age or date of birth; 5) date of recording; 6) interviewer's name.

The main grammatical categories presented are as follows:
I. Nominals (substantive, pro-substantive, adjective, pro-adjective, proper name, numeral, relative and interrogative pronouns):
1) number (sg, pl)
2) declension (15)
3) possessive suffix
II. Verbs (verbs, auxiliary verbs):
1) category for infinite form: infinitive, gerund, supine, participle
2) voice: personal, impersonal passive, personal passive
3) mood: indicative, conditional, imperative, jussive, quotative, potential
4) tempus: present, preterite, perfect, pluperfect
5) number: singular, plural
6) person: 1, 2, 3
7) affirmative or negative form
III. Uninflected words: adverbs, pro-adverbs, auxiliaries (of compound predicate), postpositions, prepositions, interrogatives, discourse markers, onomatopoetic words, conjunctions, negations, comparative words.

We have used a special program called Mark (written by Karlis Goba) to facilitate tagging and to avoid mistakes. The tagged text is in xml-format. In October 2003, about 60 000 words were morphologically tagged. Currently an Internet based search engine for the use of tagged texts is being developed.

1.4. Studies of dialect vocabulary using the corpus

A diagnostic study of the most frequent vocabulary of Estonian dialects has been carried out up to now (see also Lindström, Lonn, Mets, Pajusalu, Teras, Veismann, Velsker, Viikberg 2001). In this study one hundred most frequent words of three geographically and linguistically distinct dialects were compared. These dialects were the Western Estonian dialect, the North-Eastern dialect and the South-Eastern Võru dialect. Among the frequent words adverbs, conjunctions and discourse particles were the most numerous groups. It is usual that some words function in the text at the same time as discourse particles and adverbs, or discourse particles and conjunctions. The occurrence of such words in texts is relatively high and was of particular interest.

It appeared that the majority of these particles are common, however, for all the three dialects. From 24 analyzed particles 15 occurred in each dialect. Their phonological form may be quite different but they have the same stem (e.g. Võru *iks* ~ *i̯ks*, Western and North-Eastern *ikke* 'still; certainly, surely'; Võru *ka* ~ *kahh*, Western *koa* ~ *kaa*, North-Eastern *kaa* 'also'). Five particles occurred in two dialects and were missing in one dialect.

Only four particles were present only in one dialect among 100 of the most frequent words. These four particles are as follows. Firstly, *näet* 'to see', originally 'you (sg.) see' in Võru, which indicates that it is a verb form that has grammaticalized as a particle in South-Eastern Estonian. In the Western dialect the word was in the 269th place by its frequency and in the 2074th place in the North-Eastern dialect data. According to these text frequencies it is possible to suggest that the form is acquiring the status

of a particle in the Western dialect as well but is not used in this meaning in the North-Eastern dialect. Secondly, *vot* 'oh' in Võru that was in the 242nd place in the North-Eastern dialect and is missing in the data of the Western dialect. The same particle is typical of Russian and therefore has been mentioned among Russian loans in these eastern dialects (Must 2000 : 481). Thirdly, *ninda* 'so' that was common for the North-Eastern dialect but occurred only once in Võru and the Western dialect. And fourthly, *naa* 'so' that was characteristic of the Western dialect but did not occur in the other dialects.

We can conclude that such analysis of frequent words shows local developments as well as dialect and language contacts. Additionally, it is possible to detect the age and way of spreading of loan words. An apparent loan particle from Russian *no/nu* 'so what' occurs in all the three dialects, but was rare in the Western dialect and frequent in the eastern dialects. *Vot* 'oh' was unknown in the Western dialect and rare in the North-Eastern dialect. Also, the discourse marker *a* 'but' occurred only in the South-Eastern Võru dialect, and did not appear in the North-Eastern and Western dialect data. Thus, the South-Eastern dialect is most deeply influenced by Russian according to the use of particles; further the different stages of contacts became evident.

In Estonian dialectometrics in the 1980s and 1990s, the lexical relationships between Estonian dialects were calculated only according to the occurrence of words in the dialect (see Murumets 1982—1983; Krikmann, Pajusalu 2000) but in the future it will also be possible to count the occurrence frequencies on the basis of the data of this corpus.

1.5. Availability of the corpus

It is possible to use the sound recordings and transliterated texts in Finno-Ugric transcriptions in the form of Word-files and in simplified transcription of txt-files. At the moment there is not yet an open Internet access to the corpus but it is possible to access the database with a personal user name and password. In order to obtain these, please contact the corpus manager Liina Lindström by e-mail: liina@murre.ut.ee.

2. Basic characteristics of Estonian dialect vowel systems

The corpus has so far been used mostly for phonological studies. Recently, a statistical survey of Estonian vowel systems was carried out, the results of which will be presented here. In this study we will calculate vowel frequencies for stressed and unstressed syllables in all the Estonian main dialects in order to detect general changes in the development of vowel systems. Many, although not all of them, are related to changes in vowel harmony.

2.1. Establishing the phonemic status of the vowels

The first task is the specification of the phonemic status of the vowels. In Estonian written language, traditionally a distinction is made between nine monophthongs (see Table 5).

Table 5

Vowels of Standard Estonian

Vowel	*a*	*o*	*u*	*e*	*i*	*õ*	*ä*	*ö*	*ü*
High	–	–	+	–	+	(+)	–	–	+
Low	+	–	–	–	–	–	+	–	–
Rounded	–	+	+	–	–	–	–	+	+
Back	+	+	+	–	–	(+)	–	–	–
Front	–	–	–	+	+	–	+	+	+

The most problematic sound in the phonological description of Estonian monophthongs is *õ*. In traditional descriptions of the language and in textbooks, *õ* is treated as a mid-high unrounded vowel, i.e. similar to schwa (see e.g. Ariste 1984 : 74). Still, according to its phonetic characteristics *õ* can be classified rather as a back vowel (Eek, Meister 1994 : 409 ff), and it also has characteristics of a high vowel (see also Viitso 1981 : 68).

All short monophthongs occur in Estonian written language in the stressed first syllable; unstressed syllables contain only the so called primary vowels *a*, *o*, *u*, *e* and *i*, while *o* occurs only in the unstressed syllables in newer loan words and names. Estonian dialects, on the other hand, contain several specific monophthongs such as for example the long open *ɛ* in the stressed first syllable in West-Saaremaa, which is between *ä* and *e* in its quality, and the open *ɔ̈* instead of *õ* (see ibid.). In the non-initial syllables, the Western dialects of South Estonian have a reduced central vowel but the Eastern dialects, a high central vowel.

A common characteristic of Estonian dialects is the coarticulatory fronting or backing of vowels, as well as a certain raising or lowering of the sounds. In the following statistical analysis we have replaced the less common specific vowels with their nearest phonemes in the vowel system: *ɛ* > *e*, *ə* > *e* and *ɔ̈* > *ö*. As the only exception, we have retained the South Estonian high unrounded central vowel *i̱* (i.e. back *i*; for the description of the vowel see Viitso 1990 : 163; Parve 2000; Teras 2003). Thus, in the following, we differentiate maximally between ten vowel phonemes.

All short monophthongs have their long Standard Estonian counterparts which occur only in the primarily stressed syllables of a word. This applies also as a rule to dialects although in some dialects, long vowels can also occur in unstressed syllables. Estonian long monophthongs have been interpreted as a sequence of two short vowels (see Hint 1997 : 42—43) or as separate monophthongs (Viitso 1981). In our analysis we have first treated long monophthongs as a sequence of two vowels and then as one monophthong. Long monophthongs with the secondary quality change have been treated as the same with their closest phonetic equivalents; e.g. the overlong raised equivalents of the South Estonian long mid-high vowels have been grouped together with respective high vowels (e.g. *ū̱* > *uu*).

Estonian is rich in sequences of two different vowels; the standard contains at least 26 different types (see Viitso 1981 : 64—67; Hint 1998 : 113—115). In addition to the old diphthongs which end in an *i* or *u* there are several newer diphthongs that have developed as a result of diphthongization of long vowels, and sequences of vowels that have appeared due to the loss of a consonant. In some instances, also triphthongs are possible. The situation in Estonian dialects is extremely varied and therefore their

Number of occurrence and percentage of vowels in different Estonian dialects
(Long monophthongs are grouped together with short monophthongs)

Table 6

	a		*e*		*i*		*o*		*u*		*õ*		*ä*		*ö*		*ü*		*j*	
Mid	16878	28.04%	12660	21.03%	14482	24.06%	4640	7.71%	5286	8.78%	1602	2.66%	2850	4.73%	186	0.31%	1613	2.68%	0	0.00%
Western	13571	27.69%	10209	20.83%	10588	21.61%	3947	8.05%	4648	9.49%	1349	2.75%	3361	6.86%	212	0.43%	1118	2.28%	0	0.00%
Insular	14064	26.77%	10757	20.47%	11923	22.69%	4768	9.07%	5444	10.36%	4	0.01%	2958	5.63%	1282	2.44%	1342	2.55%	0	0.00%
Eastern	5888	25.88%	4676	20.55%	5469	24.04%	1122	4.93%	2071	9.10%	1209	5.31%	1660	7.30%	80	0.35%	576	2.53%	0	0.00%
Coastal	14005	24.43%	12519	21.84%	14053	24.52%	5181	9.04%	6274	10.95%	357	0.62%	3047	5.32%	273	0.48%	1610	2.81%	0	0.00%
North-Eastern	11240	23.90%	9036	19.21%	12666	26.93%	2778	5.91%	4460	9.48%	2367	5.03%	3251	6.91%	167	0.36%	1070	2.27%	0	0.00%
Tartu	17347	22.89%	11561	15.25%	17597	23.22%	6549	8.64%	7701	10.16%	4949	6.53%	7041	9.29%	110	0.15%	2617	3.45%	315	0.42%
Mulgi	7392	22.25%	7119	21.43%	6599	19.86%	2340	7.04%	3608	10.86%	1745	5.25%	3198	9.62%	43	0.13%	1182	3.56%	1	0.00%
Võru	11289	24.06%	6033	12.86%	10077	21.47%	4458	9.50%	5020	10.70%	3638	7.75%	4298	9.16%	78	0.17%	1646	3.51%	388	0.83%
Setu	11654	24.40%	5179	10.84%	10267	21.50%	4455	9.33%	4361	9.13%	4807	10.07%	4808	10.07%	74	0.15%	1565	3.28%	585	1.23%

Number of occurrence and percentage of vowels in different Estonian dialects
(Long monophthongs are counted as two vowels)

Table 7

	a		*e*		*i*		*o*		*u*		*õ*		*ä*		*ö*		*ü*		*j*	
Mid	17804	27.44%	13291	20.48%	16166	24.92%	5025	7.74%	5862	9.03%	1656	2.55%	3159	4.87%	259	0.40%	1660	2.56%	0	0.00%
Western	14593	26.93%	11397	21.03%	11822	21.82%	4568	8.43%	5224	9.64%	1411	2.60%	3712	6.85%	319	0.59%	1137	2.10%	0	0.00%
Insular	15345	26.25%	11977	20.49%	13413	22.95%	5316	9.09%	5841	9.99%	5	0.01%	3729	6.38%	1467	2.51%	1360	2.33%	0	0.00%
Eastern	6227	25.43%	4931	20.14%	6059	24.74%	1276	5.21%	2260	9.23%	1217	4.97%	1770	7.23%	126	0.51%	623	2.54%	0	0.00%
Coastal	15379	24.87%	12623	20.41%	15955	25.80%	5300	8.57%	6759	10.93%	359	0.58%	3420	5.53%	275	0.44%	1770	2.86%	0	0.00%
North-Eastern	12090	23.90%	9076	17.94%	14335	28.34%	2888	5.71%	4800	9.49%	2381	4.71%	3672	7.26%	168	0.33%	1174	2.32%	0	0.00%
Tartu	18916	22.68%	12035	14.43%	19393	23.25%	6996	8.39%	8956	10.74%	5249	6.29%	8346	10.01%	175	0.21%	3012	3.61%	324	0.39%
Mulgi	8028	22.31%	7236	20.11%	7405	20.58%	2471	6.87%	3951	10.98%	1767	4.91%	3709	10.31%	74	0.21%	1334	3.71%	1	0.00%
Võru	12167	23.15%	6312	12.01%	11446	21.78%	4819	9.17%	6413	12.20%	3768	7.17%	5161	9.82%	103	0.20%	1961	3.73%	407	0.77%
Setu	12587	23.92%	5536	10.52%	11257	21.39%	4771	9.07%	5273	10.02%	5085	9.66%	5602	10.65%	108	0.21%	1808	3.44%	594	1.13%

Number of occurrence and percentage of long vowels in different Estonian dialects

Table 8

	aa		*ee*		*ii*		*oo*		*uu*		*õõ*		*ää*		*öö*		*üü*		*jj*	
Mid	926	1.54%	631	1.05%	1684	2.80%	385	0.64%	576	0.96%	54	0.09%	309	0.51%	73	0.12%	47	0.08%	0	0.00%
Western	1022	2.09%	1188	2.42%	1234	2.52%	621	1.27%	576	1.18%	62	0.13%	351	0.72%	107	0.22%	19	0.04%	0	0.00%
Insular	1281	2.44%	1220	2.32%	1490	2.84%	548	1.04%	397	0.76%	1	0.00%	771	1.47%	185	0.35%	18	0.03%	0	0.00%
Eastern	339	1.49%	255	1.12%	590	2.59%	154	0.68%	189	0.83%	8	0.04%	110	0.48%	46	0.20%	47	0.21%	0	0.00%
Coastal	1374	2.40%	104	0.18%	1902	3.32%	119	0.21%	485	0.85%	2	0.00%	373	0.65%	2	0.00%	160	0.28%	0	0.00%
North-Eastern	850	1.81%	40	0.09%	1669	3.55%	110	0.23%	340	0.72%	14	0.03%	421	0.90%	1	0.00%	104	0.22%	0	0.00%
Tartu	1569	2.07%	474	0.63%	1796	2.37%	447	0.59%	1255	1.66%	300	0.40%	1305	1.72%	65	0.09%	395	0.52%	9	0.01%
Mulgi	636	1.91%	117	0.35%	806	2.43%	131	0.39%	343	1.03%	22	0.07%	511	1.54%	31	0.09%	152	0.46%	0	0.00%
Võru	878	1.87%	279	0.59%	1369	2.92%	361	0.77%	1393	2.97%	130	0.28%	863	1.84%	25	0.05%	315	0.67%	19	0.04%
Setu	933	1.95%	357	0.75%	990	2.07%	316	0.66%	912	1.91%	278	0.58%	794	1.66%	34	0.07%	243	0.51%	9	0.02%

treatment is beyond the scope of the present article. In the following statistical analysis, all sequences of vowels are separated into monophthongs.

Tables 6 and 7 give an overview of the occurrence of vowels in Estonian dialects. The treatment of long monophthongs as a unit of one or two vowels has a very small effect on the results. The number of long vowels is on the whole relatively small, the most frequent long vowel being *ii* (see table 8).

In all Estonian dialects the most common vowels are *a*, *i* and *e*. In the South Estonian dialect of Mulgi *e* is more common than *i*. The percentage of *a* ranges between 28 in North Estonian Mid dialect and 22 in the Mulgi dialect. The most frequent occurrence of *a* in the central dialects coincides with the most restricted occurrence of the *ä*-harmony in the same area, i.e. *a* can occur also in non-initial syllables of the words containing front vowels (for a detailed account see the following treatment of front and back vowels). The smallest percentage of *a* in South Estonian Mulgi dialect can be explained by the reduction of low vowels and their replacement by their mid-high unrounded equivalents starting from the third syllable (*a*, *ä* > *ə*, *e*). This explains also the frequent occurrence of *e* in the Mulgi dialect.

The percentage of *i* ranges between 27 and 20 whereas in the North-Eastern dialect group and in the Tartu dialect *i* is even slightly more common than *a*. There is, however, no easy explanation for the higher percentage of *i* in the North-Eastern dialect group as even *e* is common in these dialects. But noteworthy is the relatively stable occurrence of the historic *i* also in the diphthongs ending in an *i* and the vowel sequences that have developed due to the loss of a consonant (see also the treatment of high vowels). In the Tartu dialect, on the other hand, there is a low percentage of *a*. The reasons for this are similar to those for the Mulgi dialect.

The third most common vowel in all dialects is *e* (except Mulgi where it is on the second place). At the same time, the percentage of *e* fluctuates more between dialects than that of *a* and *i*, being for instance 21.8% in the Coastal dialect and only 10.8% in the Setu dialect. In the North Estonian and North-Eastern coastal dialects, the occurrence of *e* is even (between 21.8% and 19.2%) whereas the situation is different in the South Estonian dialects where the percentages in dialects are more uneven: 21.4% in Mulgi, 15.3% in Tartu, 12.8% in Võru and even less in the Setu dialect. In Mulgi, the frequent occurrence of *e* is linked to the extensive change in non-initial syllables: *a*, *ä* > *e*. But in other parts of South Estonia, it is its less frequent occurrence that is connected to non-initial syllables. These dialects have *õ*-harmony which means that the equivalent of *e* in non-initial syllables of the words containing back vowels is *õ*. Therefore *õ* occurs in Setu almost as much as *e* (10%) and is in non-initial syllables even more common than *e* (see the following treatment of *õ*).

The three vowels with average percentage of occurrence in Estonian dialects are *u*, *o* and *ä*. By its occurrence, *u* is the fourth most common in all the Estonian dialects except Setu where it is surpassed by *õ*. The percentage of *u* is relatively even being highest in the Coastal dialect (11%), and lowest in the Mid dialect (8.8%). There is no explicit reason for these small differences.

The fifth vowel by its occurrence is *o* in six dialects, and *ä* in four dialects. This statistics means that *o* is the primary vowel that has the most restricted occurrence in Estonian dialects. There are large differences in its

occurrence: 9.5% in Võru and only half as much (4.9%) in the Eastern dialect. The less frequent occurrence of *o* in several dialects is linked to the change of *o* into *u* in non-initial syllables, and in the Eastern dialect, to the unrounding of the *o* in the first syllable, i.e. the change *o* > *õ*. The percentage of *ä* is high in all South Estonian dialects (10.1—9.2%), considerably lower in North Estonian dialects (from 7.3% in the Eastern dialect to 4.7% in the Central dialect) and in the North-Eastern Coastal dialect group (5.3% in the Coastal dialect and 6.9% in the North-Eastern dialect). The high frequence of *ä* in the dialects of South Estonia and its low percentage in North Estonian dialects is above all caused by the larger productivity of the *ä*-harmony in South Estonia. Still, this should not be the reason for the low percentage of *ä* in the North-Eastern coastal dialects because the *ä*-harmony occurs there as well (cf Wiik 1988 : 82). It is possible, however, that here the reason lies in the diphthongization of the long *ää* into *ia* in the North-Eastern dialects.

The vowels with the most restricted occurrence in Estonian dialects are *õ*, *ü*, *ö* and *ɟ* (i.e. back *i*). The percentage of *õ* differs largely in different dialects ranging from the Setu 10.1% to the Insular 0.01%. Therefore, the occurrence of this vowel will be treated separately in the following subsection of the article. The vowel *ü* does not have a high rate but at the same time it appears relatively evenly in all dialects. Its highest percentage (3.3—3.6%) is in South Estonia where there is the *ü*-harmony and lowest in the North-Eastern dialect (2.3%). The vowel *ö* is most common in the Insular dialect (2.3%) where the central vowel *õ* has undergone rounding and turned into *ö*. In other North Estonian dialects and in the North-Eastern dialect group the percentage of *ö* is 0.3—0.5%. This figure is even smaller in the South Estonian dialects (0.13—0.17%).

2.2. *õ* in Estonian dialects

One well-known difference between the Estonian and Finnish vowel systems is the occurrence of unrounded central vowel *õ* in Estonian. But *õ* is not equally common for all Estonian dialects (see figure 2). We can see that *õ* is more atypical for the Insular and Coastal dialects and most frequent in the Eastern and South Estonian dialects.

In the case of Insular and Coastal dialects, texts from only those areas were analyzed where *õ* does not occur as a rule. In the Insular dialectal area, *õ* has undergone rounding and changed into *ö*, and in the Coastal dialects, similarly to Northern Finnic dialects, *õ* has never occurred. The fact that the dialect corpus for these dialects contains any occurrences of *õ* at all points to the beginning of levelling of these dialects with Standard Estonian.

It is, however, noteworthy that the percentage of *õ* increases gradually in the Estonian dialectal area from the North West to the South East. The percentage of *õ* is also relatively small in the North Estonian Western and Mid dialects that are the historic foundations of Standard Estonian. The percentage of *õ* is similar in the Eastern and North-Eastern dialects and in the Mulgi dialect in the Western part of South Estonia, and this in spite of the fact that none of these dialects have the *õ*-harmony and that *õ* only occurs in the first syllable.

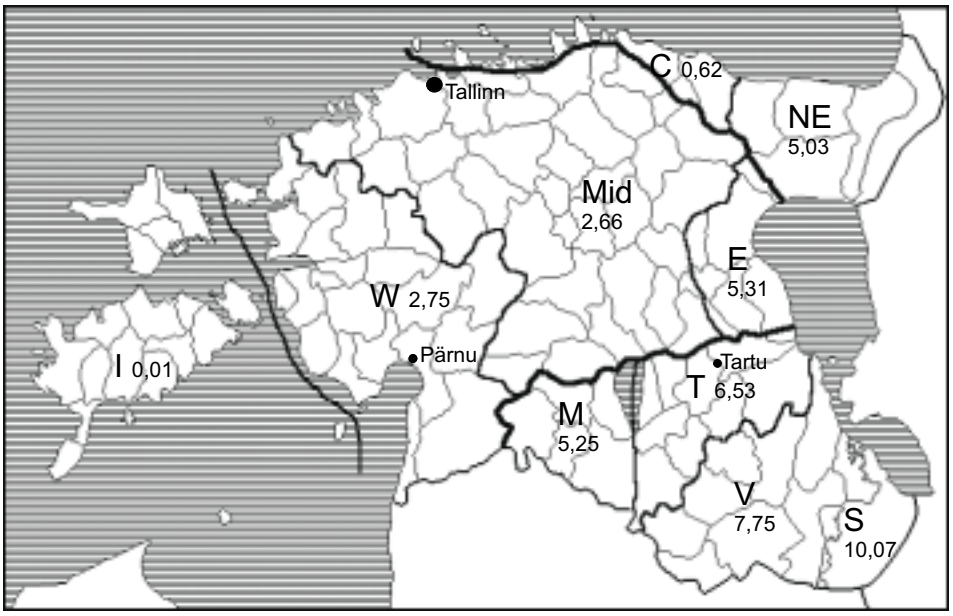

Figure 2. õ **in Estonian dialects**

The percentage of *õ* increases sharply in the South-Eastern dialects of South Estonia. The *õ*-harmony is characteristic of the dialects of Tartu, Võru and Setu but the percentage of *õ* is larger in Võru than in Tartu and in Setu considerably larger than in Võru. Do the differences in the spread of *õ* point to a more general tendency of rounding in the western dialects of Estonian and unrounding in the eastern dialects? This question will be addressed in the following analysis.

2.3. Rounded vowels in Estonian dialects

Figure 3 presents the percentage of all rounded vowels in Estonian dialects, i.e. the overall occurrence of *u*, *o*, *ö* and *ü*. As can be seen, the hypothesis presented above is only partly valid. The percentage of rounded vowels is indeed highest in the island dialects but in addition to the expected high occurrence in these dialects, unrounded vowels are also common in South Estonian dialects where a low percentage was predicted.

The general increase in the percentage of rounded vowels in the West and the decrease in the East is valid only in the case of the North Estonian dialects where we can maintain that the change *õ* > *ö* in the Insular dialects is merely part of the more general tendency to vowel rounding. As an example of this tendency is also the rounding of the *a* in the same area, as in *sauna* 'sauna' > *souna*, in Hiiumaa *a* > *å*, as in *kaks* 'two' > *kåks* (see Tauli 1956 : 174), and *i* > *ü*, *e* > *ö* next to a labial consonant as in *mitu* 'several' > *mütu*, *levad* 'loafs of bread' > *lövad*. In non-initial syllables, the rounding occurs in the Coastal dialect: e.g. in Vaivara *tulo* < *tule* 'come (Imperative)', *tulemo* < *tuleme* 'we come'. As an example of unrounding is the change *o* > *õ* in the Eastern dialects, as in e.g. *koht* 'place' > *kõht* and *oli* 'was' > *õli*, and the change of non-initial syllables in Votic: *o* > (> *õ*) > *a*, as in *ainogo* ~ *ainago* 'the only one' (see Pajusalu 2000).

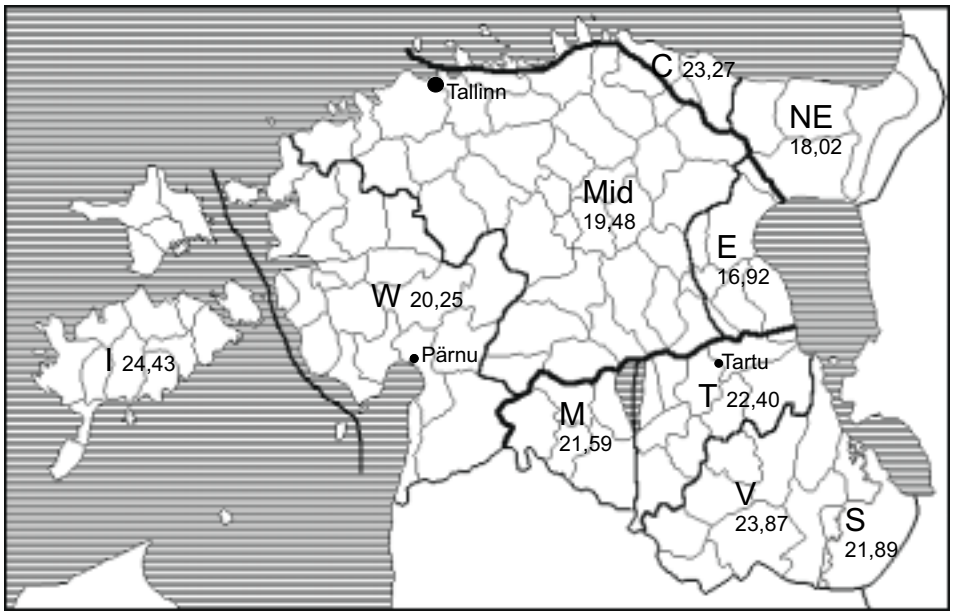

Figure 3. **Percentage of rounded vowels**

The more frequent occurrence of rounded vowels in South Estonia is most probably connected to their greater stability in non-initial syllables, as e.g. the *ü*-harmony that has prevented the change characteristic of the non-initial syllables of the North Estonian dialects: *ü* > *i*, as in ***küsü*** > ***küsi*** 'ask for'.

2.4. High and low vowels in Estonian dialects

There are also systematic differences between dialects in the percentage of high and low vowels. But here the main direction of change is not from North West to South East as in the case of the occurrence of *õ*, or from East to West as in the case of vowel rounding, but instead from South West to North East. The percentage of high vowels (*i*, *ü*, *u*, and in South Estonia also *į*) is smallest in the Western dialects and largest in the North Eastern dialects (see Figure 4). High vowels occur most of all in the North-Eastern Coastal dialect group. Among the dialects of the North Estonian dialect group, high vowels occur most in Eastern dialect although the difference is small as compared to the Mid and Insular dialects. In the South Estonian dialect group there are on average slightly more vowels than in the northern central Estonian dialects; the highest percentage is in the Tartu dialect. The lowest percentage of high vowels can be found in the South Estonian Mulgi dialect which is bordering with the Western dialect. The main reason for the differences lies most probably in the reduction and lowering of high vowels in South-Western dialects of Estonia as in non-initial syllables: *i* > *e*, e.g. *pulmalised* 'wedding guests' > *pulmalest*, *suureline* 'great' > *suurelene*, *suuresti* 'greatly' > *suureste*, (*ei*) *tohi* 'may (not)' > *tohe* but also in the lowering of high vowels in the first syllable in certain environments as in *üheksa* 'nine' > *öheksa*, *mitte* 'not' > *mette*. Such changes occur least in the North-Eastern Coastal dialect.

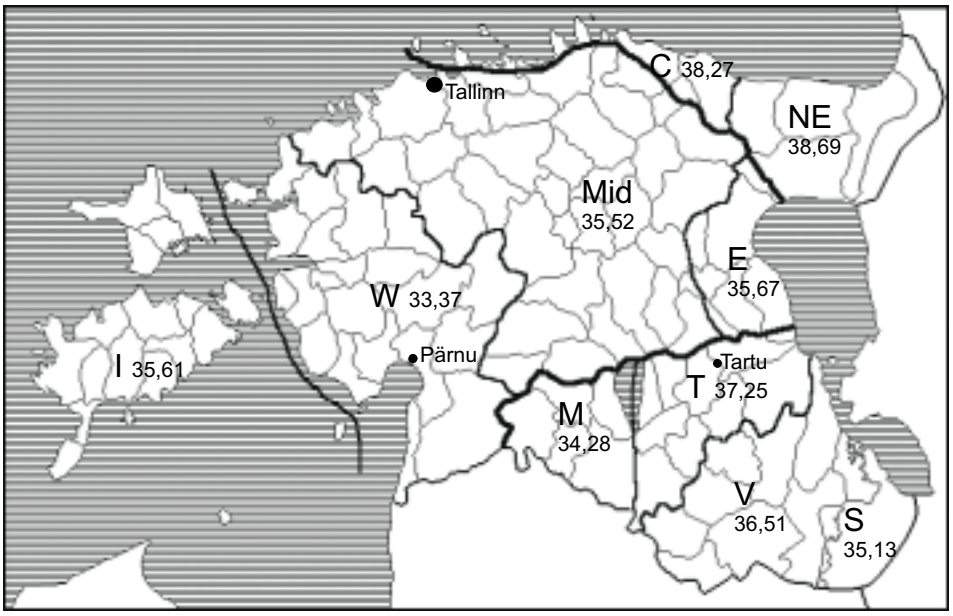

Figure 4. **Percentage of high vowels**

The percentage of low vowels (*a*, *ä*) decreases also in the direction from South West to North East if we discount the South Estonian dialects (see figure 5). The highest percentage of low vowels can be found in the Western dialect and the lowest in the dialects of the North-Eastern Coastal dialect group. Again, it is probably the Western dialect that is more innovative as here the high vowels as a rule change into mid-high, as we saw above, and mid-high vowels change into low vowels. This tendency is strongest in the southern group of the Western dialect, e.g. in non-initial syllables:

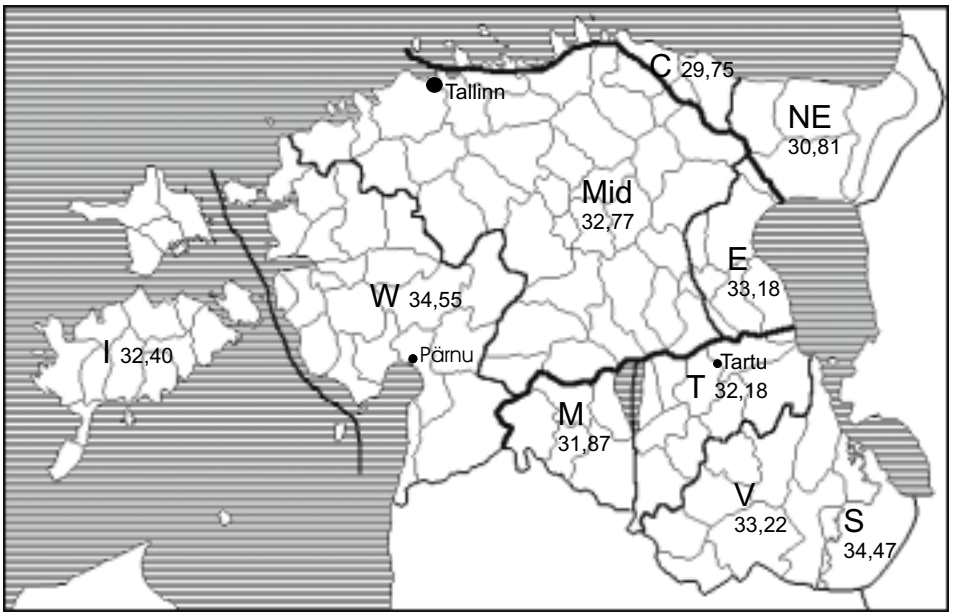

Figure 5. **Percentage of low vowels.**

e > *a* as in Tahkuranna *mere* 'of sea' > *mera*, *rehe* 'of drying barn' > *riha*. More common is the change in first syllables: *e* > *ä* e.g. *enam* 'any more' > *änam*, *vedama* 'to drag' > *vädama*, *erk* 'perky, alert' > *ärk*. In northern dialects, on the other hand, it is more common for the long *ää* and *aa* to diphthongize as in *pääseb* 'escapes' > *peaseb* ~ *piaseb*, *maa* 'land' > *moa* ~ *mua*.

South Estonian dialects, however, do not follow the general trend of change of the low vowels. Here the percentage of low vowels is smallest in the Mulgi dialect in the west, and largest in the Setu dialect in the east. This is caused by opposite changes taking place on the different edges of the dialect continuum. The western dialects of South Estonia can be characterized by the reduction of *a* and *ä* and their changing into *e* (Pajusalu 1998) as e.g. in Karksi *armastama* 'to love' > *armasteme*. On the other hand, in the eastern dialects of South Estonia, it is common for the mid-high vowels to lower in different groups (see Pajusalu 2000) as in **taloille* > *tal*(*l*)*alõ* 'to the farms'. Additionally, it is characteristic of the eastern South Estonian dialects to lower the *e* in the first syllable in some words as in *lesk* 'widow' > *läsk*, *seitse* 'seven' > *säitse*, and in Baltic loans, to have *ai* instead of *ei* as in *hain* 'hay', *saivas* 'pole', *saista* 'to stand'.

2.5. Front and back vowels in Estonian dialects

In order to clarify the distribution of front and back vowels we divided all the vowels into two groups so that the front vowels would include *i*, *e*, *ä*, *ü*, *ö* and the back vowels *a*, *o*, *u* as well as *õ* and *į* because the latter two behave similar to back vowels from the point of view of vowel harmony (although *į* can sometimes also occur in the words containing front vowels, see Parve 2000). Figure 6 presents the percentages of these back vowels.

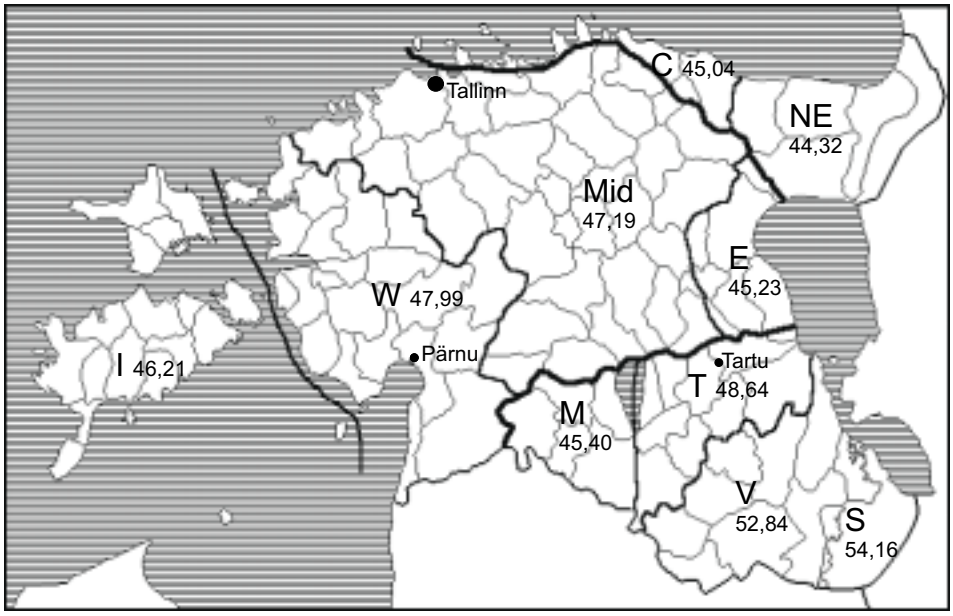

Figure 6. **Percentage of back vowels.**

It can be seen that the percentage of back vowels is relatively smallest in the North-Eastern Coastal dialect group and largest in the eastern dialects of South Estonia. This is the result of the *õ*-harmony in these dialects: where in words containing back vowels *e* has been replaced by *õ* the general percentage of back vowels is also slightly larger than that of the front vowels. The dialects where the neutral *e* is counted as a front vowel because of its phonetic characteristics exhibit a slightly larger number of front vowels. But if we discount the neutral *e* and *i* all dialects have considerably more back vowels than front vowels. It is apparent that in Estonian dialects the backness of vowels is the primary unmarked feature which is retained even after the loss of the *õ*-harmony.

2.6. Vowels in the fourth syllable

So far we have looked at the general tendencies of the vowel systems in Estonian dialects disregarding the syllables where the vowels occur. A closer look, however, shows that that there are large differences in the vowels of different syllables. We will present here as a separate example the statistics about the fourth syllable that is always unstressed (see table 9). As *į* did not occur in the fourth syllable it has been left out of the table.

Table 9

Percentage of vowels in the fourth syllable

	No.	*a*	*e*	*i*	*o*	*u*	*õ*	*ä*	*ö*	*ü*
Mid	720	23.9	56.1	15.7	–	3.9	–	0.4	–	–
Western	343	27.1	45.2	24.2	0.6	2.3	–	0.6	–	–
Insular	441	27.2	48.5	22.0	–	2.3	–	–	–	–
Eastern	326	22.7	53.4	18.1	–	4.9	–	0.9	–	–
Coastal	613	21.2	51.1	19.1	–	6.7	–	1.3	–	0.5
North-Eastern	689	26.1	43.5	21.3	–	3.3	–	5.5	–	0.2
Tartu	793	26.2	38.2	12.9	–	3.8	13.5	4.5	–	0.9
Mulgi	196	5.1	72.7	15.3	0.5	4.6	–	1.5	–	–
Võru	578	24.2	26.1	17.7	0.7	5.7	21.3	3.5	0.2	0.7
Setu	637	14.9	23.9	15.5	1.3	1.7	39.7	2.5	–	0,5

The fourth syllable which in Estonian dialects is always part of a suffix is characterized in all dialects by the restricted number of vowels. The largest number (9 vowels) appears in the South Estonian Võru dialect but as already mentioned this dialect lacks *į* in the fourth syllable and therefore does not contain the full set of vowels as well. In addition to *į*, Setu lacks *ö*, Tartu additionally *o*, and Mulgi *õ* and *ü*. The North-Eastern Coastal dialect group has six vowels in the fourth syllable: *a*, *e*, *i*, *u*, *ä* and *ü*, whereas the North Estonian dialects have only five: *a*, *e*, *i*, *u* and *ä*, and the Western dialect also has additionally *o*. Thus, the fourth syllable in all North Estonian dialects contains only primary vowels: *a*, *e*, *i* and *u*, and additionally, because of the wide-spread *ä*-harmony, *ä* in words with front vowels.

The differences in percentages are very large. The percentage of rounded vowels is the smallest, with only *u* occurring regularly. The only frequently occurring high vowel is *i* and back vowel *a*. The most common vowel in the nine dialects is the mid-high *e*, and in Setu, its back equivalent *õ* which is also widespread in Võru. If we disregard the neutral *i* and *e* it appears

that back vowels are in all dialects many times more frequent than the front vowels.

2.6. Conclusions

Several previous studies have investigated the vowel systems and sound changes of Estonian dialects but the statistical analysis presented above is the first of its kind. The results of the analysis show that several changes of single sounds are connected to more general tendencies of change in the vowel system. A more thorough study of the causes of these wide ranging changes in the vowel systems remains to be carried out in the broader context of areal linguistics. For instance, the above-described tendency to vowel rounding in West Estonia can be explained with language contacts with Swedish, and the changing of the rounded vowels into unrounded vowels in East Estonia can be due to the Slavic influence. It is apparent that the characteristic traits of the vowel system of Estonian dialects reflect often even broader characteristics of the Baltic Sea language area.

In addition to language contacts, the study of the general characteristics of the vowel systems is important from the point of view of establishing the internal rules of the systems including the markedness of the vowels. The statistical analysis showed that the primary vowels *a*, *i*, *e*, *u* and *o* are as a rule most frequent in Estonian dialects. But in four dialects *ä* is more common than *o*, and in the Setu dialect also *õ* is more frequent. In non-initial syllables of Setu, *õ* is even more common than its front equivalent *e*, which raises doubts about the markedness of *õ* in the Setu vowel harmony. The present study implies clearly that the comparison of the most general statistical characteristics of dialectal vowels enables to pinpoint several broader traits of the sound systems and their dynamics.

REFERENCES

Ariste, P. 1984, Eesti keele foneetika I, Tartu.

Eek, A., Meister, E. 1994, Eesti vokaalide sihtväärtused hääldus- ja tajuruumis. — KK, 404—413, 476—483, 548—553.

Hint, M. 1997, Eesti keele astmevahelduse ja prosoodiasüsteemi tüpoloogilised probleemid, Tallinn—Helsinki.

—— 1998, Häälikutest sõnadeni. Eesti keele häälikusüsteem üldkeeleteaduslikul taustal. Teine, ümbertöötatud trükk, Tallinn.

Krikmann, A., Pajusalu, K. 2000, Kus on keskmurde keskpunkt. — Inter dialectos nominaque. Pühendusteos Mari Mustale 11. novembril 2000, Tallinn (Eesti Keele Instituudi toimetised 7), 131—172.

Lindström, L., Lonn, V., Mets, M., Pajusalu, K., Teras, P., Veismann, A., Velsker, E., Viikberg, J. 2001, Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. — Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks, Tartu (Tartu Ülikooli eesti keele õppetooli toimetised 17), 186—211.

Murumets, S. 1982—1983, Eesti keeleala murdelisest liigendusest "Väikese murdesõnastiku" põhjal 1—2. — KK 1982, 11—17; 1983, 615—623.

Must, M. 2000, Vene laensõnad eesti murretes, Tallinn.

Pajusalu, K. 1998, Vowel Reduction in South Estonian. — LU XXXIV, 234—240.

—— 2000, Alternation of *e* and *a*, *ä* in Non-Initial Syllables in the Southern Group of the Finnic Languages. — Facing Finnic. Some Challenges to Historical and Contact Linguistics, Helsinki (Castrenianumin toimitteita 59), 156—167.

—— 2003, Estonian Dialects. — Estonian Language, Tallinn (Linguistica Uralica. Supplementary Series 1) (In press).
Pajusalu, K., Parve, M., Teras, P. 2001, On the Main Characteristics of the Prosody of South Estonian Dialects. — CIFU IX. Pars VI, 9—13.
Pajusalu, K., Hennoste, T., Niit, E., Päll, P., Viikberg, J. 2002, Eesti murded ja kohanimed, Tallinn.
Parve, M. 2000, Võru lühikeste monoftongide akustikast. — K. Pajusalu, M. Parve, P. Teras, S. Iva, Võru vokaalid I, Tartu (Tartu Ülikooli eesti keele õppetooli toimetised 13).
—— 2003, Välted lõunaeesti murretes, Tartu (Dissertationes philologiae estonicae Universitatis Tartuensis 12).
Suihkonen, P. 1998, Documentation of the Computer Corpora of Uralic Languages at the University of Helsinki. TR-2, Helsinki.
Tauli, V. 1956, Phonological Tendencies in Estonian, København.
Teras, P. 2003, Lõunaeesti vokaalisüsteem. Võru pikkade vokaalide kvaliteedi muutumine, Tartu (Dissertationes philologiae estonicae Universitatis Tartuensis 11).
Viitso, T.-R. 1981, Läänemeresoome fonoloogia küsimusi, Tallinn.
—— 1990, Vowels and Consonants in North Setu (South Estonian). — LU XXVI, 161—172.
Wiik, K. 1988, Viron vokaalisointu, Helsinki (Suomi 140).

ЛИЙНА ЛИНДСТРЕМ (Тарту), *КАРЛ ПАЮСАЛУ* (Тарту—Хельсинки)

КОРПУС ЭСТОНСКИХ ДИАЛЕКТОВ И СИСТЕМА ГЛАСНЫХ ЭСТОНСКОГО ЯЗЫКА

Первая часть статьи знакомит с электронным корпусом эстонских диалектов, над составлением которого начиная с 1998 г. работают исследователи Тартуского университета и Института эстонского языка, и с накопленным к настоящему времени опытом его использования. В основу корпуса легли старейшие звукозаписи эстонских диалектов начиная с 1930-х годов, большая часть материала — записи 1956 — 1979 гг. Корпус состоит из точно соответствующих звукозаписям текстов в финно-угорской транскрипции и их вариантов в упрощенной транскрипции, причем последние снабжены морфо-синтаксическими пометами, а потому уже теперь можно осуществлять морфологический анализ текста автоматически.

По состоянию на октябрь 2003 г. корпус включает тексты на всех эстонских диалектах, всего 456 000 слов текста. Наряду с четырьмя североэстонскими диалектами — островной, западный, центральный и восточный — отдельно представлены тексты на прибрежном и северо-восточном диалектах из северо-восточной прибрежной диалектной группы, а также южноэстонские диалекты — мульгиский, тартуский и выруский — и отдельно сету. Отбирались тексты по возможности более центральных говоров каждого диалекта.

Вторая часть статьи представляет собой статистическое сравнение эстонских диалектов относительно вокализма. Выявлены систематические различия между диалектами. Например, увеличение доли лабиальных гласных в ареале распространения эстонских диалектов последовательно наблюдается в направлении с запада на восток, а уменьшение доли иллабиального $\tilde{o}$ — с северо-запада на юго-восток. Наиболее употребительны лабиальные гласные в островных диалектах, а $\tilde{o}$ — в диалекте сету. Систематические различия в вокализме между диалектами позволяют объяснить ряд отдельных звукоизменений как составную часть более общих процессов звукоизменения, имеющих, вероятно, более широкую подоплеку. Исследование показывает, что наблюдение общих характеристик систем гласных, наряду с выявлением ареальных особенностей, имеет большое значение и для определения основных правил вокализма, в том числе немаркированных гласных.