Alphabetic Proportions in Estonian Monolingual and Bilingual Dictionaries

Enn Veldi
University of Tartu

The paper discusses alphabetic proportions in Estonian general monolingual and bilingual dictionaries with Estonian on the left-hand side. As no data about the Estonian alphabetic proportions were available, the alphabetic proportions were calculated on the basis of the corpus-based Frequency Dictionary of Standard Estonian *(Kaalep and Muischnek 2002). The findings were then used to configure the Estonian ruler in TshwaneLex dictionary compilation software-alphabetic proportions for English, Afrikaans and several Afrikan languages as well as an excellent background to this problem can be found in De Schryver 2005. Subsequently, the established proportions were used as a yardstick for comparing three monolingual and six bilingual dictionaries. The six bilingual dictionaries included four general Estonian-English dictionaries-one of them not completed as of yet but revealing potential problems-and two school dictionaries-Estonian-German and Estonian-Russian. The findings show that while alphabetic proportions have generally been followed quite successfully, some Estonian dictionaries show a tendency to be skewed-some dictionaries become more thorough towards the end of the alphabet while others show the opposite trend. The problem is more challenging for those dictionary projects that require decades for completion and where the dictionary is published in fascicles-in the Estonian lexicographic practice one has in mind the* Explanatory Dictionary of Standard Estonian*, which started publication in 1988. There can also naturally be instances where single alphabetic stretches may reveal perceived overtreatment or undertreatment. The paper also argues whether the alphabetic proportions of certain letters can vary to some extent depending on the selection of words listed under them-e.g. inclusion of large numbers of foreign and learned words beginning in* a, b, d, g , f *in Estonian dictionaries can increase the proportions of these letters as they are less typical of native words. The paper ends with the firm conviction that in recent years it has become much easier to control the progress of a lexicographic project.*

1. Introduction

Until now Estonian lexicographers have had to learn the hard way how to overcome the problem of alphabetic proportions in their dictionaries. Therefore, when Gilles-Maurice de Schryver (2005) recently focused on concurrent over-and undertreatment in dictionaries, I realized that the problem should be studied also in the Estonian context. The present paper will first establish the alphabetic proportions for Estonian dictionaries and then compare the findings with the proportions of three monolingual and six bilingual dictionaries with Estonian on the left-hand side.

2. Alphabetic proportions in Estonian

As no data about the alphabetic proportions were available in Estonian, the author decided to calculate them on the basis of the word list in the *Frequency Dictionary of Standard Estonian* (Kaalep and Muischnek 2002). This dictionary lists about 9,700 words, which were established on the basis of a corpus of one million words of newspaper texts and fiction. The analysis of the frequency list yielded the following distribution of letters across the Estonian alphabet.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
5.30	0.49	0	0.56	4.1	0.46	0.28	4.17	2.33	2.86	16.11	6.41	6.40	3.07
515	48	0	54	398	46	27	405	226	278	1565	623	622	298
O	P	R	S	Š	Z	Ž	T	U	V	Õ	Ä	Ö	Ü
2.18	9.23	4.11	8.45	0.05	0	0.01	9.71	1.15	8.65	1.08	0.59	0.13	2.09
212	897	399	821	5	0	1	943	112	840	105	57	13	203

Table 1. Distribution of the word list in the *Frequency Dictionary of Standard Estonian* (Kaalep and Muischnek 2002) across the Estonian alphabet expressed in % and number of lemmas

The findings were then also used to configure the Estonian ruler in TshwaneLex dictionary compilation software.

3. Comparison of three general monolingual dictionaries

The findings were subsequently used as a yardstick for comparing three general monolingual Estonian dictionaries. *ÕS 2006* stands for the 2006 edition of the standard Estonian *Dictionary of Correct Usage* of about 13,000 entries; *Õpilase ÕS* (2004) is a shorter dictionary of Standard Estonian for secondary-school students and has about 22,000 words, and EKSS (1988-2007, *Explanatory Dictionary of Standard Estonian*) is an unabridged seven-volume descriptive dictionary, which has been published in fascicles since 1988. It should also be mentioned that *ÕS 2006* is a treasure house of Estonian-language terminology. The publication of EKSS has not reached the final four letters of the Estonian alphabet as yet; therefore, the values of the frequency dictionary were used to calculate the prospective overall length of the dictionary. All the above-mentioned dictionaries were available to the author as print editions; for this reason the following estimates are based on page counts.

	Kaalep & Muischnek	ÕS 2006	Student ÕS	EKSS
A	5.30	6.02	3.7	3.02
B	0.49	1.08	0.91	0.35
C	0	0.06	0.03	0.001
D	0.56	1.57	1.3	0.56
E	4.1	3.07	3.36	2.09
F	0.46	1.21	0.86	0.54
G	0.28	0.88	0.68	0.35
H	4.17	3.89	4.26	2.97
I	2.33	2.02	2.26	1.36
J	2.86	2.27	2.9	1.99
K	16.11	15.61	17.75	15.06
L	6.41	5.73	6.66	5.29
M	6.40	6.51	5.27	6.17
N	3.07	3.07	2.77	3.75
O	2.18	1.45	1.18	1.49
P	9.23	11.34	8.65	12.92

R	4.11	5.06	3.78	6.05
S	8.45	8.28	8.28	10.83
Š	0.05	0.16	0.17	0.13
Z	0	0.02	0.03	0.02
Ž	0.01	0.03	0.07	0.04
T	9.71	8.77	9.12	11.09
U	1.15	1.04	1.49	1.51
V	8.65	7.48	10.1	9.06
W	0	0.005	0.03	0.002
Õ	1.08	1.24	1.45	0.94
Ä	0.59	0.47	0.68	0.50
Ö	0.13	0.15	0.17	0.11
Ü	2.09	1.57	1.82	1.79

Table 2. Alphabetic proportions in the *Frequency Dictionary of Standard Estonian* Kaalep and Muischnek and three monolingual dictionaries (*ÕS 2006*, *Õpilase ÕS*, and EKSS) expressed in %

The findings enable us to put forward several interesting hypotheses, which can be subsequently checked by a closer look at the macrostructure of the respective dictionaries. First, the higher values for the letters *a, b, d, g, f* in *ÕS 2006* in comparison with the other dictionaries may indicate a higher proportion of listed foreign and learned words in this dictionary. Second, it appears that the longest alphabetic stretches in Estonian are *k, p, s, t, v*, which require special attention on the part of the lexicographer. Also, more data are needed to find out why the proportions of *p* and *t* are higher or lower in different dictionaries. Third, the table shows that EKSS has gradually become more thorough towards the end of the alphabet. This seven-volume dictionary has been published in fascicles over twenty years at a rate of a fascicle per year. It is quite clear that the second edition of this dictionary needs balancing throughout the first part. Fourth, *Õpilase ÕS* may need some fine-tuning; one might claim that the letters *k* and *v* have been to some extent overtreated, and the letters *m, p, r, t* reveal undertreatment.

4. Comparison of six bilingual dictionaries

It was also of great interest to compare the alphabetic proportions of the frequency dictionaries with a number of bilingual dictionaries where Estonian appears on the left-hand side. The data from five of them can be found in the table below. *SAAGPAKK EST-ENG*, *TEA EST-ENG*, and *SILVET EST-ENG* represent Estonian-English dictionaries; *SAAGPAKK EST-ENG*, which is the most comprehensive dictionary in this category, includes ca. 150,000 entries. *KOOLIBRI EST-RUS* (Estonian-Russian School Dictionary, 21,000 words) and *TEA EST-GERM* (Estonian-German School Dictionary, 15,000 words) represent dictionaries for secondary school students. The data about all the above-mentioned these dictionaries were obtained by page counts.

	Kaalep & Muischnek	SAAGPAKK EST-ENG	TEA EST-ENG	SILVET EST-ENG	KOOLIBRI EST-RUS	TEA EST-GERM
A	5.30	4.15	5.61	3.75	4.03	4.89
B	0.49	0.61	0.74	0.43	0.64	1.01
C	0	0.03	0.007	0	0.01	0.03
D	0.56	0.83	1.04	0.53	0.39	1.18
E	4.1	3.13	3.93	2.53	3.77	4.49
F	0.46	0.62	0.81	0.47	0.59	0.78

G	0.28	0.47	0.41	0.33	0.39	0.54
H	4.17	4.02	4.77	3.32	4.16	5.5
I	2.33	1.67	2.52	1.69	1.83	2.74
J	2.86	2.68	3.11	2.41	3.02	3.49
K	16.11	14.99	16.02	14.85	17.52	17.42
L	6.41	6.09	5.53	6.25	6.8	7.32
M	6.40	5.54	6.38	6.27	7.68	7.67
N	3.07	3	3.14	3.46	3.13	3
O	2.18	1.5	1.83	1.79	1.63	2.32
P	9.23	10.77	9.50	9.96	10.4	9.73
R	4.11	6.25	4.85	5.11	4.41	2.82
S	8.45	10.76	9.57	10.08	7.91	7.25
Š	0.05	0.12	0.13	0.19	0.14	0.17
Z	0	0.03	0.007	0.03	0.01	0.001
Ž	0.01	0.03	0.03	0.07	0.02	0.03
T	9.71	9.24	8.13	10.68	7.91	7.15
U	1.15	1.17	1.58	1.39	1.03	1.08
V	8.65	8.45	7.32	9.53	8.03	6.39
W	0	0.1	0	0	0	0
Õ	1.08	1	1.08	1.47	0.99	0.88
Ä	0.59	0.64	0.43	0.75	0.94	0.49
Ö	0.13	0.19	0.13	0.18	0.14	0.17
Ü	2.09	1.74	1.40	2.3	2.39	1.5
X	0	0.01	0	0.01	0.001	0
Y	0	0.01	0	0	0.001	0

Table 3. Alphabetic proportions in the Frequency Dictionary of Standard Estonian Kaalep and Muischnek and five blingual dictionaries (SAAGPAKK EST-ENG, TEA EST-ENG, SILVET EST-ENG , KOOLIBRI EST-RUS, and TEA EST-GERM) expressed in %

The table enables us to notice the following. First, the percentages of the huge dictionary by the old master Saagpakk correlate surprisingly well with those in Kaalep and Muischnek. One can notice, however, that the values for *k* and *t* are slightly lower and those for *p* and *s* are somewhat higher. *TEA EST-ENG* shows lower values for *t* and *v*, which are the last two longer alphabetic stretches. *SILVET EST-ENG*, which represents the work of the other Estonian old master Johannes, is somewhat more thorough in the second half of the dictionary where many letters show somewhat higher values. *KOOLIBRI EST-RUS* has the highest value for *k* and also the values for *l* and *m* are higher while the values for *s, t* and *v* are lower. A similar tendency can be observed in *TEA EST-GERM* where the values for *k, l*, and *m* are high, but the end part of the dictionary shows a downward trend in the values for *s*, *t*, *v*, and *ü*.

Finally, there is an Estonian-English dictionary (Aule 2001, 2003), which was in 2000 advertised by the publisher as a single-volume dictionary with 50,000 entries and one thousand pages in length. However, when the dictionary reached bookshops a year later, it appeared that the published volume had about 600 pages and covered only the letters from A to J. Judging by

the published volume it was to be a five-volume dictionary at the time of completion. So one was definitely looking forward to the second volume, which came out in 2003 (367 pages) and covered the letter K (the most challenging stretch in Estonian dictionaries). The second volume was supplied with a slip from the publisher, which informed the subscribers that the project had turned out to be more extensive and thorough than originally planned, which would contribute to the better quality of the dictionary. In 2003 the publisher also stated that the dictionary would be published by volumes in 2003-2008, that the completed dictionary would have 80,000 entries taking up least 2,500 pages. At the time of writing this paper volume three is not available as yet.

A	B	C	D	E	F	G	H	I	J	K
5.30	0.49	0	0.56	4.1	0.46	0.28	4.17	2.33	2.86	16.11
3.39	0.49	0.06	0.67	3.39	0.63	0.45	4.89	3.15	3.58	14.48

Table 4. Alphabetic proportions in the *Frequency Dictionary of Standard Estonian* (Kaalep and Muischnek) and the published letters of the Estonian-English dictionary by Aule (2001, 2003) expressed in %. The percentages for the Estonian-English dictionary were calculated on the basis of the projection that the dictionary will be 2,500 pages in length at completion.

Here one can see that the end of the first volume shows some overtreatment in the letters *h*, *I*, and *j* while the published second volume (the letter *k*) reveals some undertreatment. It is interesting to follow the alphabetic proportions of the future volumes of this dictionary.

5. Conclusion

It is quite clear that the recent arrival of the ruler function as a feature of dictionary compilation software is of great help in monitoring the alphabetic proportions in the course of writing the dictionary. The findings showed that most of the studied dictionaries revealed a degree of unevenness—some dictionaries became more thorough towards the end of the alphabet while others showed the opposite trend. Consistency is a challenge for those dictionary projects that take decades to complete and where the dictionary is published in fascicles. In such cases one can at first publish a supplementary volume followed by a balanced second edition. One can also argue whether the alphabetic proportions of some letters, such as *a, b, d, g, f* in Estonian, could show variation that depends on the degree of listed foreign and learned words beginning with these letters.

References

Aule, A. (2001, 2003). *The Contemporary Estonian–English Dictionary*. Tallinn: Estonian Language Foundation.

De Schryver, G. M. (2005). "Concurrent Over- and Under-treatment in Dictionaries—The Woordeboek van die Afrikaanse Taal as a Case in Point". *International Journal of Lexicography* 1. 47–75.

[EKSS]. *Eesti kirjakeele seletussõnaraamat*. (1988–2007). Tallinn: Valgus.

Kaalep, H. J.; Muischnek, K. (2002). *Eesti kirjakeele sagedussõnastik*. Tartu.

[KOOLIBRI EST-RUS]. Leemets, H. (2005). *Eesti-vene koolisõnaraamat*. Tallinn: Koolibri.

[Õpilase ÕS]. Erelt, T.; Leemets, T. (2004). *Õpilase ÕS*. Tallinn: Eesti Keele Sihtasutus.

[ÕS 2006]. Erelt, T. et al. (2006). *Eesti Õigekeelsussõnaraamat ÕS 2006*. Tallinn: Eesti.

Keele Sihtasutus [on line]. *http://www.eki.ee/dict/gs2006* [Access date: 30 March 2008].

[SAAGPAKK EST-ENG]. Saagpakk, P. F. (1992). *Estonian-English Dictionary*. Tallinn: Koolibri.

[SILVET EST-ENG]. Silvet, J. (1980). *Estonian-English Dictionary*. 2nd edition. Tallinn: Valgus.

[TEA EST-ENG]. Mägi, R. (ed.) (2005). *Estonian-English Dictionary*. 1st edition. Tallinn: TEA.

[TEA EST-GERM]. Ellert, L. et al. (eds.) (2007). *TEA Koolisõnastik. Estnisch-Deutsch*. Tallinn: TEA.