Frequency Dictionary of Written Estonian of the 1990ies

Heiki-Jaan Kaalep

Filosoft Ltd.
Tartu, Estonia
hkaalep@filosoft.ee

Kadri Muischnek

University of Tartu
Tartu, Estonia
Kadri.Muischnek@ut.ee

Abstract

The paper describes an (almost) automatically created frequency dictionary of Estonian literary language of the 1990-ies. The dictionary is based on a 1-million word text corpus which contains fiction texts and quality nation-wide newspapers in equal proportions. Morphological analysis, disambiguation and lemmatization were done automatically, although some post-editing was necessary. The dictionary and the corpus are both available on-line.

1. Introduction

In this paper we describe the Frequency Dictionary of Literary Estonian (Kaalep, Muischnek 2002). The dictionary is based on the written texts of 1990ies; the size of the text corpus is 1 million words.

As Estonian is a heavily inflected language, the frequency counts are based on a lemmatized corpus. The texts were lemmatized, using Estyhmm (Kaalep, Vaino 2001).

In addition to the description of the creation of the frequency lists, the cumulative coverage of the corpus by various ranks of most frequent words is calculated.

The dictionary itself and the corpus that it is based on can both be downloaded from http://www.cl.ut.ee.

2. Frequency and commonness

Frequency of a word is highly correlated with its commonness. Common words tend to occur frequently in texts, and vice versa, words that are rarely met in texts may be classified as uncommon.

In order to interpret the information in a frequency dictionary, it is important to bear in mind that although related, frequency and commonness are not synonyms. E.g. "kägu" (cuckoo) is unquestionably a common Estonian word, but it is frequent only in certain text types like songs and fairy-tales, not in newspapers or literary prose (and thus it did not cross the threshold to get into this dictionary). Frequency in a single text or even a text class does not guarantee respective commonness for the word. Frequency is calculated from a certain amount of texts, and it is dependent on the type of these texts. Many words that are frequent in university-level textbooks of physics are uncommon for the language as a whole; the same is true for fairy-tales. Worse still, a frequent word may be uncommon even for the sole text class it belongs to. Words do not occur randomly in texts, but according to some theme the text is about. This in turn means that any ranking of words that is based on their frequency misrepresents their commonness ranking. In addition to frequency, one has to take into account the distribution of a word in different texts. If a word occurs in many texts, although only a few times in each, it is more common than a word that occurs equally many times, but in a small number of texts. (See also (Kilgariff 1997).)

When treating frequency as a measure of commonness, the texts we are basing our counts on should be homogenous to some extent. If a text corpus consist of texts from very different text classes (e.g. internet relay chat and legislation), then what do the frequency counts really represent?

3. Corpus

The current dictionary is based on a 1-million word text corpus of literary prose and nation-wide general interest quality newspapers. These two are both large, well-defined and homogenous text classes, being at the same time not too different from each other. Together they should represent standard

wide-spread neutral Estonian literary language.

The proportion of the text classes in the corpus is 50-50. The literary prose texts (published in 1992-1998) are taken from the Tartu University text corpus of contemporary Estonian (http://www.cl.ut.ee). The corpus consists of 2000-word excerpts; a few texts are represented by more than one excerpt. The newspaper texts (published in 1995-1999) are taken partly also from http://www.cl.ut.ee, and partly from internet archives of the newspapers, to create a more homogenous and balanced corpus. Newspapers have been included in total, not by 2000-word excerpts.

When interpreting the frequency counts of this dictionary as a measure of general commonness of Estonian words, one has to be cautious, in view of the small size of the text corpus (only 1 million words) and the fact that the corpus does not represent many important text classes, and most notably, speech transcripts. (Compare it with "Word Frequencies in Written and Spoken English" (Leech et al, 2001), which is based on a 100-million word British National Corpus.) However, the only frequency dictionary of Estonian thus far has been the theoretically well-based "Eesti keele sagedussõnastik" (Frequency Dictionary of Estonian) (Kaasik et al, 1976, 1977) which was based on a mere 100,000 word text corpus of literary prose from 1960ies, representing only one text class – the author's speech.

The newspaper corpus contains 510,200 word forms, and the literary prose corpus contains 496,800 word forms, making the corpus of 1,007,000 word forms, including numbers, proper names, abbreviations and acronyms. Without these, the corpus contains 908,400 word forms that have been the basis of the frequency counts.

4. Estyhmm

For finding the lemmas (base forms) of the words, we first processed the corpus with estyhmm (Kaalep, Vaino 2000), a morphological analyzer and bigram HMM-disambiguator for Estonian, the output of which contained for every word its lemma and part of speech (word class). After that we counted the occurrences of the lemmas in newspapers, literary prose and in the corpus as a whole.

Automatic processing of texts was not entirely error-free. A major deficiency was the need to classify the plural forms (except plural nominative) of pronouns "see" (this) and "tema" (he, she, it) – "nende", "neid", "nendes" etc. – which are homonymous and can be differentiated on semantic grounds only. This was done manually and thus the frequency counts of "see" and "tema" are correct. A few other lemmas required post-editing too:

Many Estonian proper names are homonymous with a common noun in some declined form. This poses problems for their automatic treatment, as it is hard to tell whether a word form is a common noun or a proper name (e.g. given names like "Kalju" (rock), "Laine" (wave), and especially compound family and place names), and a dictionary should not include proper names. We tried to eliminate such errors by hand-validating the frequency lists. For example, we deleted "Mustamägi" (black mountain, a place-name) and re-counted words like "liiv" (sand, a frequent family name) and "mari (berry, a frequent given name).

If the texts have been analyzed automatically and we know that the program is not 100% correct, it is important to know how trustworthy the result – frequency dictionary – is.

To evaluate this, we checked two versions of similar texts, one containing base forms, found by the program, and the other containing base forms, found by a human annotator. The most common mistake was that the program treated a proper name as a common noun. Of all the running words, 2% had been incorrectly classified as common nouns, e.g. "Kõuts" (tomcat), "Väli" (field). The impact of this error was diminished by our choice not to include words that are not met both in newspapers and literary prose: many proper names occur in only one text class, or even one single text, so they were left out as uncommon ones. We also browsed all the words of the dictionary and in case of a suspiciously high frequency count, checked the occurrences manually in the texts, and adjusted the counts.

Besides the incorrect discrimination between proper names and common nouns, the automatic analyses contained a wrong lemma form for 0.75% of the words. In reality the error count should be smaller still: the

dictionary contains only summarized counts, and instances of an incorrectly classified word balance each other to sum up to figures that may be closer to truth. In any case, an error level of 0.75% is comparable to imprecision, resulting from the choice of texts as the basis of the dictionary.

5. Lemmatization

When reading and using this dictionary, one must remember that it contains frequency counts of words, not senses. E.g. the frequency of the verb "tulema" (come; have to) summarizes the frequencies of both senses.

The component words of expressions (like idiomatic verbs or phrasal verbs) are counted as independent words. Thus the expression "aru saama" (understand) is split into "aru" (sense; homonymous with "aru", grassland) and "saama" (get).

Concatenated compounds are very frequent in Estonian texts. We made no attempts to split them into simplex words. We treated any string of characters, separated by white space, as a word form.

We discarded proper names, abbreviations, acronyms and numbers from the frequency dictionary.

One should not make too far-fetching conclusions from frequencies of single words. Borrowing from John Sinclair, a British linguist and lexicographer, we may say that the meaning of even single words cannot always be inferred from the words themselves in isolation, as the meaning is dependent on the context, expression. So although we have a very frequent word in our dictionary, "aeg" (time), mostly it does not denote an ontological or abstract category, but simply occurs in expressions like "samal ajal" (at the same time), "viimasel ajal" (lately), "kogu aeg" (all the time), "pikka aega" (for a long time). By way of comparison, it is noteworthy that "time" is the most frequent word in the Frequency Dictionary of Finnish (Saukkonen et al 1979).

Every lemma in the dictionary belongs to a word class: noun (S), adjective (A), numeral (N), verb (V), pronoun (P) or indeclinable word (D); indeclinable words are adpositions (pre- and postpositions), adverbs, conjunctions and interjections. A lemma may belong to more than one word class at a time (as explained below). Two lemmas – "oma" (one's; one's own; ours; about) and "pool" (half; side; spool; at) have even 4 word class tags; others have less.

6. Dictionary Size and Corpus Coverage

This frequency dictionary actually shows only the tip of an iceberg: the whole number of lemmas in the corpus was 60,000, 32,000 of which occurred only once.

How frequent and widely distributed a word should be to get included in a frequency dictionary, is an issue in itself. The aim of this dictionary is to list common Estonian words, so the criterion is very strict: every word must occur both in newspapers and literary prose. If it is missing from either of them, it is not common enough to be included.

Only 14,500 lemmas were met both in the newspapers and literary prose, and of these, 9,700 occurred at least 5 times, which was the threshold for getting included in this frequency dictionary.

23,500 lemmas were found solely in the newspapers and 22,000 solely in literary prose.

E.g. "puuraidur" (lumberjack) occurred 50 times in prose, but never in newspapers. Newspapers, in their turn, contained "omavalitsus" (local authority) 209 times, without it ever having been mentioned in literary prose.

The following table shows the cumulative coverage of the text corpus by successive frequency ranks of the lemmas, sorted in a descending order of frequencies. The figures in the first two columns are rounded.

Table 1. Cumulative coverage of the text corpus by lemmas

First … words	Corpus coverage in %	Frequency count at least
10	19,3	6194
20	24,6	4032
50	33,1	1797
100	40,7	1034
250	51,3	452
500	60,2	229
1000	69,0	115
1500	74,0	72
2000	77,2	52
3000	81,5	30
5000	86,0	15
10000	90,3	5

We see that the 250 most frequent words cover over 50% of the text corpus, and that 10,000 most frequent words cover about 90% of the texts.

The following table shows the cumulative coverage of the text corpus by successive frequency ranks of the word forms, sorted in a descending order of frequencies. The figures in the first two columns are rounded.

Table 1. Cumulative coverage of the text corpus by word forms

First … word forms	Corpus coverage in %	Frequency count at least
10	13,0	5329
20	17,2	2961
50	23,5	1445
100	29,4	863
250	38,2	373
500	45,3	187
1000	52,4	95
1500	56,7	65
2000	59,7	50
3000	64,2	33
5000	69,7	20
10000	76,9	10
20000	83,8	5
33000	88.8	3

We can see that we need 33,000 different word forms to achieve the 90% text coverage, i.e. 3 times as many as lemmas for the same text coverage.

7. Conclusion

In this paper we have presented the Frequency Dictionary of Estonian, the corpus that the dictionary was based on and the principles for lemmatization and lemma selection.

We assume that the work we have described here has at least twofold results: the dictionary itself and the tested procedure that can be repeated using a different or simply a much bigger corpus.

The frequency lists we have created can be used for solving a whole range of problems, in theoretical linguistics as well as in applied and computational linguistics. It has already been used in a computer program that checks the complexity of various school textbooks.

8. Bibliographical References

Kaalep, H-J., Muischnek, K. 2002. *Eesti kirjakeele sagedussõnastik* (Frequency Dictionary of the Estonian Literary Language). TÜ kirjastus, Tartu.

Kaalep, H-J., Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9-16, Tartu.

Kaasik, Ü., Tuldava, J., Viilup, A., Ääremaa, K. 1976. Eesti keele ilukirjandusproosa autorikõne sõnavormide sagedussõnastik. (Frequency dictionary of word forms of author's speech in Estonian literary prose) *Keelestatistika 1. TRÜ toimetised 377*, Tartu, pp. 107-153.

Kaasik, Ü. Tuldava, J., Villup, A., Ääremaa, K. 1977. Eesti tänapäeva ilukirjandusproosa autorikõne lekseemide sagedussõnastik. (Frequency dictionary of lexemes of author's speech in Estonian literary prose) *Keelestatistika 2. TRÜ toimetised, 413*, Tartu, pp. 5-140.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. In *International Journal of Lexicography* 10, pp. 135-155

Leech, G., Rayson, P., Wilson, A. 2001. *Word Frequencies in Written and Spoken English*. Longman, Pearson Education.

Saukkonen, P., Haipus, M., Niemikorpi, A., Sulkala, H. 1979. *Suomen kielen taajuussanasto*. (A frequency dictionary of Finnish) Werner Söderström osakeühtiö. Porvoo - Helsinki - Juva.

.