THE CORPORA OF ESTONIAN AT THE UNIVERSITY OF TARTU: THE CURRENT SITUATION

Heiki-Jaan Kaalep, Kadri Muischnek
University of Tartu, Estonia

Abstract

This paper gives an overview of the corpus-related work done at the University of Tartu so far and describes an ongoing project – compiling a big corpus of written Estonian containing approximately 100 million words. The previously collected corpora of standard written Estonian at the University of Tartu are well-balanced and representative, but a little too small for the studies of statistically not so frequent phenomena in language, not to speak of the needs of language technology. The corpus under compilation right now, called the Mixed Corpus of Estonian, is planned as an open monitor corpus, but will also contain a more balanced subcorpus.

In addition to these corpora of standard written Estonian, the paper gives a very brief overview of the Corpus of Estonian Dialects, The Corpus of Spoken Estonian and the Corpus of Old Literary Estonian and discusses some special annotated corpora in more detail, namely the morphologically annotated corpus and the Estonian-English parallel corpus of legislative texts.

Keywords: Estonian language corpora, corpus linguistics, corpus compilation, corpus annotation

1. Introduction

In the recent years the focus of corpora-related work at the University of Tartu has been on building a big corpus of Estonian, consisting of at least 100 million words. A really big language corpus is essential for computational linguistics and for more theoretical branches of Estonian linguistics as well.

The easiest way to obtain written texts is to collect the texts that are already in an electronic form. We started from the texts available via Internet. Newspaper text is the dominating text type there, but one can find also legal texts, scientific texts, etc. We are trying to avoid manual work (downloading, converting, tagging) as much as possible. So we use special computer programs that do all this. Our final goal is to have a text that has been annotated up to the level of sentences, i.e. the headings, paragraphs, sentences and highlighted words/phrases are marked.

The work has been financed by the Ministry of Education via a national program "Eesti keel ja rahvuskultuur" (Estonian Language and National Culture).

2. The previous corpora at the University of Tartu

The history of corpus linguistics at the University of Tartu began in the first half of the nineties, when the 1-million word Corpus of Written Estonian was compiled. The work followed the well-known principles of Brown and Lancaster-Oslo/Bergen corpora: the corpus was divided into ten text classes that were designed to represent the whole written (edited and printed) culture from the years 1983-1988, the central year being 1985. This is a balanced sample corpus, each text excerpt containing maximally 2000 words. In addition to this, nine balanced subcorpora were compiled, one for each decade of the period 1890-1990, except for the 1980ies that were covered by the first corpus. These subcorpora contain about 300-400 thousand words each and they contain only two text classes: press and fiction - the largest text classes in Estonian culture and the only ones that exist through the 20th century in Estonian. The criteria of selection were the same as in the first corpus (for longer overview about selection of the fiction and press, see (Hennoste et al 1998)). Altogether this Thread of Corpora contains a little more than five million words. It gives a good overview of the development of the Estonian language during the 20th century, and has been in extensive use especially by students. But still it remains too small for studies addressing linguistic phenomena of lower frequency.

To make the picture complete, we should give a short overview of some other corpora being compiled at the University of Tartu as well.

The Corpus of Spoken Estonian (http://sys130.psych.ut.ee/~linds/) contains about 600 thousand words of transcribed speech, mostly everyday and institutional conversations. For more detailed description the reader is referred to (Hennoste et al 2000 and Hennoste et al 2001).

Closely related to the previous corpus is the Estonian Dialogue Corpus (http://www.cs.ut.ee/~koit/Dialoog/EDiC). It contains three different types of dialogues: 1) spoken human-human dialogues, 2) written human-computer simulated interactions (using Wizard of Oz method), 3) human-computer interactions. The dialogues have been annotated for dialogue acts and communicative strategies. The reader will find a more detailed description of the dialogue corpus in (Koit 2002 and Koit 2003).

The Corpus of Estonian Dialects (http://www.murre.ut.ee/) contains about 600 thousand words at the moment. The corpus contains texts in phonetic transcription, in simplified transcription, as well as morphologically annotated texts. One can use the corpus via Internet user interface. The reader will find a detailed description of the Corpus of Estonian Dialects in (Lindström and Pajusalu 2003, Lindström et al 2001).

The Corpus of Old Literary Estonian (http://www.murre.ut.ee/vakkur/) contains over 700 thousand words and covers the period from the year 1224 up to the end of the 18th century. The corpus can be accessed via Internet user interface. For detailed description of the corpus the reader is referred to (Kingissepp et al 2004).

3. The Mixed Corpus of Estonian

3.1. The problem of representativeness and text classes

The ideal corpus of written language should represent all the text types that exist in the written culture of that language and the proportion of every text class in the whole corpus should correspond to the proportion of this text class in the whole body of written texts in a certain period. This is of course difficult to accomplish. In the history of corpus linguistics the well-known examples of well-balanced corpora are the Brown Corpus and the Lancaster-Oslo/Bergen Corpus. They contain only one million words each but have still remained valuable language resources. The British National Corpus that was compiled in the first half of the nineties "is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written." as the homepage of BNC describes it (http://www.natcorp.ox.ac.uk/).

Most of the corpora of various languages are not balanced, mostly due to two reasons: 1) building an unbalanced corpus demands less resources and less time and 2) deciding on the sublanguages and text classes that should be included and their proportions in the corpus is a difficult task and would doubtlessly be an object of severe criticism by the future users of the corpus.

This new Mixed Corpus of Estonian is planned as an open monitor corpus, meaning that it will not be necessarily balanced or representative. However, a smaller balanced subcorpus is also being compiled. It is described in more detail in section 3.3.

The new corpus contains only whole texts, no text samples. While the Thread of Corpora described in the previous section contained only texts initially written in Estonian, e.g. no translations, then for this corpus we collect also the translated texts. But there is one exception – we have decided not to include translated fiction or if included, it must be kept strictly separate from the fiction written in Estonian. The reason for this is the extremely bad translation quality of some fiction texts, especially in the text class of so-called commercial fiction (detective stories, love stories, etc.). For example the studies of word order based on these texts would show Estonian word order being very similar to that of English (but that is not the case, of course).

Like the Thread of Corpora, this new corpus will also contain no drama or poetry; it will contain only the written texts meant for reading – i.e. pre-planned and post-edited language usage mostly. But we will include more spontaneous and informal written speech, namely the language of newsgroups, internet forums and chatrooms.

The process of planning and collecting the corpus has revealed the fact that some text classes are underrepresented or totally absent in Estonian. For example it is quite difficult to find a scientific article in physics written in Estonian.

3. 2.The Current Situation

At the moment the Mixed Corpus contains the following subcorpora:
1) daily 'Postimees', 33 mio words
2) weekly 'Eesti Ekspress', 7,5 mio words
3) weekly 'Maaleht', 4,3 mio words
4) Estonian fiction, 4,2 mio words
5) PhD dissertations 500,000 words
6) popular science journal 'Horisont', 260,000 words
7) academic journal 'Akadeemia', 7 mio words
8) transcripts of Estonian Parliament (Riigikogu), 13 mio words
9) weekly "Kroonika", 600,000 words
10) Estonian legislative documents, 1,8 mio words
11) Estonian translations of EU legislation, 9,6 mio words

The newspaper text class is clearly overrepresented. We have two reasons for that. The first one is pragmatic: converting the newspaper texts into the corpus format gave us maximal amount of words with minimal effort. But we also find the language used in newspapers being the closest to the so-called "general or standard Estonian".

3.3. The Balanced Corpus

To enable some comparative studies of the three (main) text classes of written Estonian, we have planned a balanced subcorpus within the Mixed Corpus. This will contain newspaper, fiction and scientific texts, five million words each. The newspaper part of it has been completed already, the fiction part we hope to complete soon, but the collecting and converting of the scientific texts still needs a considerable effort.

3. 4. Annotation

The general mark-up follows the TEI Guidelines (http://www.tei-c.org/Guidelines2/). The non-ascii characters are represented as SGML entities.

The division of the texts into paragraphs follows the original files. The headings and authors have been tagged. The text inside paragraphs has been processed by a program called estyhmm; as a result, the punctuation marks are separated from wordforms by a space (except those punctuation marks that are an integral part of the token, e.g. an abbreviation or an ordinal number) and the sentences are tagged with <s> and </s>. Every file starts with a header <teiHeader> documenting the file contents, size, used tags etc.

As for the markup of the initial structure of the text, the daily “Postimees” could serve as a nearly ideal example: the corpus has been divided into single newspaper issues, subdivisions of newspapers and newspaper articles, each of them being tagged as a division of separate level. As a result of this every sentence in the user interface can be linked to a source description giving the article, the author of the article, the subdivision and the newspaper issue were this particular sentence was printed.

4. Special subcorpora

In addition to these text collections our group has prepared some subcorpora with extra levels of annotation.

4.1. Morphologically disambiguated corpus

In this corpus (http://test.cl.ut.ee/korpused/morfkorpus/index.html.en) the text has been automatically analysed by a program called estmorf (Kaalep, Vaino 2000) and subsequently manually disambiguated by two persons; and the third person has compared the result and made the necessary corrections.

The disambiguated texts belong to the following text classes:

Text class	number of tokens
Fiction (Estonian authors)	104 000
G. Orwell's "1984"	75 500
Newspaper texts	111 000
Legal texts	121 000
Texts from a popular science journal "Horisont"	98 000
Reference texts	4 000
Altogether	513 000

The word-forms have been analyzed one by one, except for some multi-word proper names like New York. The result of the analysis contains:

1) segmentation of the word into morphemes (stems and affixes)
2) lemmatization of the rightmost stem
3) syntactic word-class tag
4) morphological categories

Ca 0,3% of the analyses can be debatable due to the ambiguousness of the borders between word classes or wrong because of human mistakes.

For more detailed description of the morphologically disambiguated corpus the reader is referred to (Kaalep et al 2000 or Muischnek and Vider 2005).

4.2. The Treebank of Estonian

The reader can learn about the Treebank (a corpus annotated for the phrase structure) of Estonian from (Uibo and Bick this volume).

4.3. The Estonian-English Parallel Corpus of Legislative Texts

This corpus contains:

1) Estonian-English parallel texts, 1.7 million tokens in Estonian, 2.9 million tokens in English.

2) English-Estonian parallel texts, 3.3 million tokens in Estonian, 4.9 million tokens in English.

The texts originate from Estonian Legal Language Centre (http://www.legaltext.ee) on April 30, 2002. The aligned versions are based on the TEI P3 compatible versions of the same files from the Mixed Corpus of Estonian.

The texts have been sentence-aligned. The items of lists are treated as equal to sentences. The Estonian and English sentences may be in 1-1, 1-2 or 2-1 alignments. There are no other alignments (like 1-0, 0-1, 2-2 etc) in this corpus. They were either not found or they were left aside as they would be hard to use in future work, the aim of which is to find parallel multi-word units. The aligning was done using the Vanilla aligner (http://nl.ijs.si/telri/Vanilla/). It is a language independent aligner, based on the algorithm from (Gale, W. A. and Church, K. W. 1993).

5. User interface

The Mixed Corpus of Estonian and the Corpus of Written Estonian could be used via Internet interface at http://test.cl.ut.ee/korpused/kasutajaliides/index.html.en

All the texts are divided into sentences, so one always gets a full sentence as an answer to her/his query. It is also possible to ask for up to five preceeding and following sentences and so get more contextual information. It is only possible to seek for a wordform (or string including regular expressions) as these corpora have not (yet) been morphologically annotated. At the moment all the texts in the user interface are represented as plain text, all tags removed.

The morphologically annotated corpus described in 4.1. can be accessed via its own interface at http://test.cl.ut.ee/korpused/morfliides/index.html.en

6. Conclusion

This paper gave an overview of the various corpora of Estonian at the University of Tartu. The main focus was on the corpora of the standard written Estonian, especially on the open monitor corpus called the Mixed Corpus. The latter is under construction at the moment, so its main compiling principles were discussed. Some special corpora with more detailed annotation (i.e. morphological annotation) were also described.

References

Gale, W. A.; Church, K. W. 1993. Program for aligning sentences in bilingual corpora. *Computational Linguistics* 19, 75-102.

Hennoste, Tiit; Roosmaa, Tiit; Saluveer, Madis 1998. Structure and usage of the Tartu University Corpus of Written Estonian. *International Journal of Corpus Linguistics.* 3 (2), 279-304.

Hennoste, Tiit; Lindström, Liina; Rääbis, Andriela; Toomet, Piret; Vellerind, Riina 2000. Eesti suulise kõne korpus ja mõnede allkeelte võrdluse katse. In: Tiit. Hennoste (ed) *Arvutuslingvistikalt inimesele*, *Tartu Ülikooli üldkeelteaduse õppetooli toimetised 1*. Tartu, pp 245-283.

Hennoste, Tiit, Lindström, Liina, Rääbis, Andriela, Toomet, Piret, Vellerind, Riina. Tartu University Corpus of Spoken Estonian. In: *Congressus Nonus Fenno-Ugristarum Pars V*. Tartu 2001, pp 345-351.

Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela; Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. *Keel ja Kirjandus vol 9*, pp. 623-633.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2001. Complete Morphological Analysis in the Linguist’s Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp 9-16, Tartu.

Kingisepp, Valve-Liivi, Prillop, Külli, Habicht, Külli 2004. Eesti vana kirjakeele korpus: mis tehtud, mis teoksil. *Keel ja Kirjandus vol 4*, pp 272-283

Koit, Mare 2002. Kommunikativnye strategii v informacionno-spravochnom dialoge (na materiale estonskogo korpusa dialogov). In: *Proc. DIALOG-2002, 6-11 June 2002 b. Vol. 2*, Moskva, Nauka, 283-290.

Koit, Mare 2003. Märgendatud dialoogikorpus kui keeleressurss. In: *Toimiv keel I. Töid rakenduslingvistika alalt. Eesti Keele Instituudi toimetised 12*. Tallinn: Eesti keele Sihtasutus, 119-136.

Lindström, Liina; Pajusalu, Karl 2003. Corpus of Estonian Dialects and the Estonian vowel system. *Linguistica Uralica 4*, pp 241-257

Lindström, Liina; Lonn, Varje; Mets, Mari; Pajusalu, Karl; Teras, Pire; Veismann, Ann; Velsker, Eva; Viikberg, Jüri 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. In: *Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks. TÜ eesti keele õppetooli toimetised 17*, pp 186-211.

Muischnek, Kadri; Vider, Kadri 2005. Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis. To apper in: *Töid Rakenduslingvistika alalt.*

Uibo, Heli; Bick, Eckhard 2005. Treebank-based research and e-learning of Estonian syntax (*this volume*).

Viks, Ülle 1992.Väike vormisõnastik I. Sissejuhatus & grammatika; II. Sõnastik & lisad. Tallinn.