Tiit Hennoste, Heiki-Jaan Kaalep, Kadri Muischnek,
Leho Paldre, Tarmo Vaino
Tartu

The Tartu University Corpus of Estonian Literary Language

We are going to describe the Corpus of Estonian Literary Language of the 20th Century (CELL) and its internet-based interface (http://www.cl.ut.ee/corpusb), as well as some of the problems we encountered during the creation of the corpus, the tagging process and building the interface. This article consists of two parts. At first we describe the Corpus of Estonian Literary Language and the principles of the selection of the texts it contains. Here we also discuss some general problems of corpus linguistics. In the second part of the article we describe the corpus interface - a Unix command window.

Introduction

Text corpora have become a usual resource of the linguistic evidence for the linguists as well as for the computational linguists, not to mention the lexicographers.

Not only the texts of the big languages like english, german or french, but also of the languages with much smaller amount of speakers and thus also texts are collected into corpora. Nevertheless, perhaps the statement that John Sinclair made in the year 1991 about corpora and corpus-based work still holds: The results are only as good as the corpus, and we are at a very primitive stage of understanding the character of corpora and the relation between decisions on the constitution of the corpus and information about the language derived from the corpus (Sinclair 1991:9).

What kind of corpora do we need?

A linguist is interested in both diachronic and synchronic corpora. A synchronic corpus might mean a collection of texts from a period of hundreds of years, but some language changes might be investigated over much shorter periods. For example, the language used in the Estonian newspapers has undergone a quite remarkable change during 1980-1999.

On the other hand, when compiling a corpus for language technology one should focus on collecting as modern language as possible. If aiming at constructing language technology tools that can analyse the present-day language, one should avoid using the texts containing lexical units and constructions that differ too much from it, i.e. from the language our tools are built to analyse. So using newspaper texts of the eighties from the CELL or Soviet propaganda texts from the same subcorpus for creating lexicons or training the algorithms for the 21st century would rather impair the performance of the tools.

An ideal corpus should contain as much different text classes as possible. The so-called classical corpora - Brown and Lancaster/Oslo-Bergen, as well as parts of the CELL, compiled according to their principles - are divided carefully into several text categories. It would be a corpus linguist's dream to have a large, balanced and representative corpus, but due to the lack of time and/or resources, we have to contend either with a small well-balanced corpus, or to collect more texts and allow the corpus to be unbalanced. Moreover, we lack knowledge about the possible future uses of electronic texts: different users are interested in different aspects. Not including electronically available texts only because they would tilt our corpus would be a disservice for the potential users.

What have we really got?

At the moment we have both carefully balanced sub-corpora, as well as un-balanced ones. On the web page of our research group the following corpora are available.

1. Balanced corpora
1.1. The basic CELL, containing texts from 1983-1987
1.2. Synchronic corpora of the 20th century
1.2.1. Fiction and newspapers from 1890-1899
1.2.2. Fiction and newspapers from 1900-1910
1.2.3. Fiction and newspapers from 1911-1920
1.2.4. Fiction and newspapers froms 1935-1939
1.2.5. Fiction and newspapers from 1945-1954
1.2.6. Fiction and newspapers from 1966-1970
1.2.7. Fiction and newspapers from 1971-1975
1.2.8. Fiction and newspapers from 1988-1998
2. Un-balanced corpora

2.1. Newspapers from 1999, collected during the ELAN project (http://solaris3.ids-mannheim.de/elan/) - 365 000 words.
2.2. The Estonian version of G. Orwell's novel '1984' - 75 000 words.

The basic CELL - the texts from the years 1983-1987 were all typed in from the keyboard. In the same way most of the texts in the additional corpora were converted into the electronic form, smaller amounts of the texts were converted into the electronic form by optical scanning. The subcorpus of the 1990ies was either scanned or obtained in electronic form directly from the publishers. During a EU Copernicus project ELAN ca 500,000 words of newspaper texts were collected from Internet and converted into the TEI-format (http://www.tei-c.org) automatically. '1984' is very different from the rest of the corpora: it represents a whole book (not excerpts), it is a translation, and it has been manually tagged for morpho-syntactic phenomena.

During 1994-1997 we created a 50-million word electronic archive of the news produced by the local news agencies BNS (Baltic News Service) and ETA (Estonian Telegraphy Agenture). But we must admit we do not use the texts obtained in that way in our corpus. An important reason for that is that we do not have rights to make the texts freely available to researchers outside our own institution. Converting electronic texts into a proper corpus requires considerable time and effort, and we are not willing to spare any if the result will be confined only to a handful of people. Another reason we haven't included the news agency texts in our corpus is that at the news agencies, only some 10 persons do write or translate the news and so we would find ourselves investigating the idiolects of those 10 people.

The selection criteria.

The balanced sub-corpora, 8 altogether, follow the same principles as those underlying the classical Brown and Lancaster-Oslo/Bergen (LOB) corpora. We could divide those subcorpora into three groups:

1. Basic CELL or the subcorpus from 1983–87. It contains 1 million words and it was the first we compiled. The texts were chosen from the same text classes as in LOB: press, science, fiction, religion, skills&trades&hobbies, popular lore, belles lettres&biography, and miscellaneous. Two text classes were added: encyclopaedia and soviet propaganda. Of course, there are differences in the content and capacity of the classes between the CELL and LOB (a longer overview about selection of the texts in basic-CELL can be found in (Hennoste 1996, Hennoste et al 1998)).

Table 3
Texts in the basic CELL

Categories	**Years**	**Texts**	**Words**	**%**
ABC **Press**	1985	519	176017	17,2
D **Religion**	1984–6	4	8011	0,8
E **Skills, trades, and hobbies**			75410	7,4
Books	1984–6	20	39572	
Periodicals	1984–6	45	35838	
F **Popular lore**			164218	16,1
Books	1984–6	49	150024	
Periodicals	1985	37	14194	
G **Belles lettres, biography**			90661	8,9
Books	1985, 5, 7	16	32017	
Periodicals	1984, 5	36	58644	
I **Miscellaneous**				
Documents	1984–6	8	12427	1,2
J **Learned**			155448	15,2
Books	1984–6	49	96235	
Periodicals	1985	37	59213	
KLMNPR **Fiction**			255416	25,0
Books	1984–7	93	192667	
Periodicals	1984–7	35	62749	
S **Encyclopaedias**	1984, 5, 8	11	22769	2,2
T **Ideology**			60256	5,9
Books	1984–6	14	28638	
Periodicals	1985	16	31618	
TOTAL	1984–88	989	1020645	100

The other balanced subcorpora of CELL are about 300-400 thousand words each and they contain only two text classes now: press and fiction. Those are 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 basic CELL (longer overview about selection of the fiction and press, see Hennoste, Muischnek 2000).

Table 4
Texts in the other subcorpora

	KLMNPR **Fiction**	ABC **Press**
Year	Words	Words
1890–1899	155 000	193 000
1900–1910	64 500	171 500
1911–1920	247 000	182 500
1935–1939	252 000	117 000
1945–1954	66 000	–
1948–1952	–	242 400
1966–1970	132 000	201 000
1971–1975	257 100	168 500
1984–1987 (Basic)	255 416	176 017
1988–1998	611 000	384 800

For a text to be selected in the corpus, the text had to be:
-produced and circulated in Estonia (not abroad),
-prose (not poetry),
-from public/formal situations,
-mother tongue text (not translations),
-written, edited and printed,
-designated from educated adult authors to adult readers
-from the first edition/printing only (not second editions or second printings)

The language of those texts corresponds roughly to literary standard Estonian.

The number of extracts in each text-class is roughly in accordance with the distribution of those text classes in Estonian culture.

Selection criteria for the fiction were the following:
1. From each book of Estonian prose, one 2000-word extract was chosen
2. In addition, some 2000-word extracts were chosen from the main literature magazines ("Looming" and "Noorus" in the 1960ies, "Vikerkaar" in the 1990ies).

Selection criteria for the press were the following:

1. All the newspapers that were published in the corresponding period were divided between a few predefined classes. Texts were chosen so that they would represent all the classes.
-two papers that had been published throughout the whole century - "Postimees" and "Sakala" - were included in every subcorpus
-by the frequency/rate of the publishing the papers were divided into dailies (5-7 times a week), weeklies and those that had 2-3 issues weekly,
-by the circulation area the newspapers were divided into nation-wide, newspapers of the large towns (Tallinn and Tartu), and local papers,
-by the content and target audience the newspapers were divided into general and specialized papers (eg sport, culture), quality papers, and tabloids.
-the newspapers from the beginning of the 20-century were divided additionally to the left and rightwing papers, because their language was very different. There was no such a division in the other periods

2. In the LOB, an additional selection was made by the genres of the newspaper texts (news, reportage, columns and so on). In contrast, we selected not single texts from the newspapers, but whole issues, because it was impossible to divide the newspaper stories into the classical genres in Soviet era and at the beginning of the 20th century. Thus, all the original (not translated) editorial texts were included from the newspapers.

3. The number of the issues from each newspaper is in accordance with the distribution of the newspaper in the corresponding period.

Corpus annotation

All the texts are tagged at least up to the level of sentences, i.e. the headings, paragraphs, sentences and highlighted words/phrases are marked. Some sub-corpora are tagged according to TEI guidelines (http://www.tei-c.org): the basic CELL, '1984' and those newspaper texts since 1985 that are also included in the ELAN corpus (http://solaris3.ids-mannheim.de/elan/)

In the basic CELL, two thirds of the newspaper texts and half of the fiction texts were tagged for sub-sentence elements like proper names, lists, foreign-language material, abbreviations, numerals, date and time. But these phenomena were tagged manually and so the mark-up is inconsistent.

Corpus annotation is a very labour-consuming process. Thus, it would be rational to annotate the corpus as much as necessary, but as little as possible. The information about the source and encoding of the text (represented usually in the text header) is inevitable - what is the use of data without any information about its source or reliability? The same is valid for the annotation of text structure - the retrieval and concordance programs must be able to distinguish between the different sentences and words in the text. We are convinced that the texts must be divided at least into chapters, paragraphs, sentences and words. But would it be reasonable to store the texts also in a morphologically analysed and disambiguated form, or should the corpus interface enable us to perform the morphological analysis and disambiguation on the run, during answering a corpus query, remains an open question at the moment.

The annotation should be homogeneous throughout the whole corpus. Our experience with using the basic CELL (with all its heavily tagged parts) has convinced us that a corpus is really tagged up to the level of its least-tagged subpart. So it is reasonable to complete one stage of mark-up in the whole corpus and then proceed to tagging the next level. Otherwise, the effort that has been made to tag a part of the corpus in a finer level is simply lost.

Unix as the Interface of an On-line Text Corpus

A corpus should be usable via Internet, to allow maximum access to it. When devising our corpus query interface, we wanted to avoid complications resulting from "making a special corpus interface for linguists" which would most likely mean a slightly unconventional syntax, partly lengthy and partly missing documentation, and buggy routines. Instead, we decided to provide the users a Unix command line window, in addition to a simple grep-like query for less complicated search.

The increasing sizes of language corpora and higher demands on their processing leads to finding new ways to make corpora accessible. It is not conceivable any more that a linguist would turn to a computer specialist every time he has a more complicated task than just a simple lookup of a word.

This has motivated us to making corpora widely available via Internet and providing them with intelligent tools to process them. The Unix interface we describe has been implemented on the Corpus of Estonian Written Language that is being built by the University of Tartu.

Description of interface

The Corpus of Written Estonian (CELL) is internally represented as a set of lines: one line is one sentence. We offer the user a Unix interface via WWW (http://www.cl.ut.ee/cgi-bin/unix_sj_en.cgi) so that relevant Unix commands can be entered for queries over (parts of) the corpus. Commands are run on complete sentences (i.e. lines) and can be used singularly or piped into queries that are more complex. How powerful the interface is depends very much on one's knowledge of Unix. In principle, there are endless combinations of commands.

This kind of interface was prompted by the need of linguists to have a friendly and powerful access to the corpus. Our first attempt was to provide a grep-like query for simple concordances. It can be used both over texts with first level tags (sentences) and for morphologically tagged texts. Similarly, one can use regular expressions and/or search for morphological tags in CELL.

The first solution was sufficient for simple concordances but did not allow further refinements of results nor queries that are more complicated. We thought it wise not to invent any further *ad hoc* query language of our own that would need thorough documentation and differ from conventional languages mostly by annoying modifications in the names of commands.

Traditional Unix facilities seemed a reasonable solution. Unix commands have become standard and they are sufficient for most corpus queries, so no additional features or pre-processing is necessary. In principle adding language specific tools, e.g. morphological analyser or disambiguator, is only a technical problem (that we have not yet implemented). In addition, one can easily add any filters in order to conform to the copyright requirements.

For the sake of security only a choice of Unix commands has been made available for corpus research: *cut, egrep, grep, head, join, paste, rev, sed, sort, tail, tr, uniq, wc*. Each of them is linked to an appropriate man-page so there is no need for extra documentation. Constraint that no output can be saved to the WWW server should guarantee abuses of the interface.

Our simpler query is quite similar to several other interfaces. For example, IMS Corpus Query

Processor developed by University of Stuttgart (Cf. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/) allows special attributes and uses regular expressions in queries. British National Corpus (Cf. http://thetis.bl.uk/lookup.html) offers a more limited search possibility. TransSearch (Cf. http://www-rali.iro.umontreal.ca/TransSearch/ TS-simple-uen.cgi) is even simpler but suggests parallel context in another language as output. We have not dealt with queries on parallel-translated texts, as CELL is a monolingual corpus.

We have not met a Unix interface implemented as a query language elsewhere. The interface we provide makes no limitations on the amount of corpus processed and downloaded. It includes most of the facilities of the query languages mentioned above in addition to the extra possibilities.

At present, it runs on corpus of several million words that actually form subcorpora of texts from different decades of this century. The contemporary texts have been morphologically tagged. The older ones are plain texts because the morphology has slightly changed over times and the morphological analyser needs to be adjusted. In principle, the power of our interface depends on how much the corpus has been tagged.

Our corpus interface has been a tool for Estonian language research for both university students and researchers. Mostly the grep-like concordance page is used for simple queries in order to get example sentences for various purposes. This can be concluded by the fact that the number of hits per visitor to this page is relatively small.

Most of the users, however, have presumably been students of our own university who have needed it for their assignments. Nevertheless it has provided very useful for scientists with specific tasks who carry out research on actual language usage, compile mono- and multilingual dictionaries and explore language change.

Our experience with students has been that without having any prior knowledge of Unix one can obtain basic skills and use Unix as a query language after four hours of tutoring. In reality, this has meant 2-3 practical classes within a computer class oriented to students of linguistics who had had only a general experience of using a computer. Explaining regular expressions took one third of the time. Creating a frequency list based on CELL was the final task of one course. This assignment can be found at http://www.cl.ut.ee/en/unix_examples.html Most students were able to complete it.

True, the main complaint that we have heard as feedback is that Unix is too complicated to be used by a usual linguist. We think that any interface would need a certain amount of instruction in order to enter as complex queries as can be done with Unix commands. It needs further testing whether learning Unix is faster than other query languages or not.

Another feedback has been that in case of Unix queries one easily gets an awful lot of sentences that takes time to download. So far processing time and memory usage has not been a problem to our server. And we have advised our students to start with a *head*-command while experimenting.

Evaluation of the corpus interface

There are several issues with our interface that might cause problems. Although they do not directly concern the tool itself, they are to be taken into account when implementing and developing such an interface.

When corpora grow larger we come to the question of processor load and slow connections. We cannot compare the speed of our searches to other tools because it depends very much on external things. For the same reason we have not thought it reasonable to provide the user with feedback about how processing is going on and how long the current process is likely to take. It is, however, possible to cancel the job and return to the previous stage without disturbing the system.

It is not possible to set any kind of standard limit to the amount of output in the case of Unix queries. In case of searches the output could well be limited to number of lines (sentences). But when we add *sort* and *uniq* facilities then this kind of limitation becomes ridiculous.

We have met some occasional error messages «Document contains no data» when 15 students have tried to perform the same query simultaneously during a computer class. In fact, we are not concerned with the computational power and bandwidth: the hardware is developing so quickly to keep in pace with video and sound processing and transporting, that text processing and transporting are really no problem.

As a prerequisite for such a corpus interface, the copyright restrictions on the texts of the corpus have to be minimal. One can achieve it by creating the corpus by collecting excerpts, like in the LOB (Cf. http://www.hit.uib.no/icame/lobman/lob-cont.html) and Brown (Cf. http://www.hit.uib.no/icame/brown/bcm.html) corpora. This is what we have done also with the balanced sub-corpora.

Copyright problem concerns chiefly the output. This can be limited by various means depending on whether limitations are set on the amount of text or on the range of users. Neither option is especially appealing, though.

Another prerequisite for using this corpus is some amount of knowledge of Unix commands. We believe that with the help of example queries, however, this obstacle can be overcome. Unix users can have very different levels of command. Most frequent queries can easily be learned.

Currently the user gets no original Unix error messages if he has made a mistake in the command line. The only error message we have implemented prompts the user when he has used an unknown (or unallowed) command.

At present the internal format of the corpus is sentence-per-line. In the morphologically annotated part, the analyses are on the same line with the words, delimited by special symbols; the whole sentence is one single line. As a further development the internal format might be made more explicit so that the user can take more advantage of it.

HTML characters have been used for special symbols in order not to make queries dependent on any particular code page. This makes entering the text somewhat clumsier. A solution would be to let the user choose how to enter these characters.

We haven't made any effort to link the corpus with some dictionary, or the texts with parallel texts. As Estonian disambiguators develop and become more exact the quality of output will increase.

We believe that our interface has several advantages due to the fact that the query language in Unix.

With Unix as interface one can make very complicated queries, if necessary. This flexibility is confined to the fact that one cannot save anything to a file nor can one combine two or more files as input. This would make *join* command much more powerful.

Piping commands gets very useful in refining output when the initial output includes systematically unnecessary information. Finally the output can be personally designed so that important items are easily recognizable (turning keywords into bold, highlighting them with special symbols etc.) And there is always the possibility to save the output to one's own computer and continue analysing with other tools.

The number of available commands does not have to be limited to our choice. The corpus interface provider could add new tools which the users can exploit in a command pipe. In linguistic research a morphological analyser and a disambiguator would be marvellous tools to use. Those would then need extra documentation, of course.

We have not modified Unix in any way. The query language (Unix commands in a pipe) is a well-known standard - that has not changed since it was developed in 1976 - with ample documentation. In practice this can be one of the most releaving arguments for setting up such a tool: no need to create and update user manuals.

It needs no deep skill to put up an interface like ours. A practical hint is to use GNU commands instead of the ordinary ones. Sentences in a corpus can be extremely long, and GNU commands support lines of infinite length.

Query languages usually assume some kind of advanced structure of input. This can be so and is an advantage in case of Unix, but not necessarily so. Unix tools are powerful enough on plain untagged text. Further tagging adds linguistic relevance to queries but the engine remains the same.

Some prospects for the future

After our successful experiment with the newspaper corpus in the course of the ELAN project, where we found that by far the easiest way to collect a corpus is from Internet, we have started to plan a new corpus project. We intend to collect various Estonian texts directly from publishers and from Internet, to achieve a corpus of 100 million words.

In the context of an automatic collection of large amount of texts, we intend to pay special attention to collecting texts as different from each other as possible, to compensate for the lack of a carefully balanced choice of texts. We would include everything from quality newspapers to parliament debates, from tabloids to school textbooks, from legal and bureaucratic texts to Internet chat-room transcripts.

The main difficulties in compiling such a corpus are of course the copyright problems. Especially they tend to complicate the collection of fiction texts. It seems also to be complicated to get an exact overview of the texts existing on the web. So we don't have any clear comprehension about the amount of scientific, popular science texts or teaching materials available via Internet.

At present we have started to program downloading and conversion tools for some newspapers and for the archive of Estonian parliament debates. The parliament debates starting from the year 1995 are available from Internet without any restrictions and we have signed agreements with the publishers of two daily newspapers - "Postimees" and "Eesti Ekspress". A corpus of parliament debates from 1995 - 2000 would contain 10 million words; a corpus of the aforementioned two newspapers from 1995 - 2000 would contain 20 million words.

We estimate that one could collect 30 million more words of newspaper and magazine texts from Internet, but in order to do that, we must first get the permissions from the publishers to use them in our corpus.

Bureaucratic and legal texts (laws, court decisions etc) also present a text class, available in large amounts electronically. We intend to include those texts in our corpus as well.

A serious gap in our collections is the lack of big parallel corpora (estonian-english, estonian - russian). Nevertheless, we assume that the texts suitable for compiling this kind of corpus do exist and that they are just waiting to be picked up.

As to the interface, we plan to add linguistic tools, like a morphological analyser and a disambiguator, to the basic Unix ones.

References

Hennoste, Tiit 1996. Tartu University Corpus of Written Estonian: A Survey of the Structure of Texts and Principles of Selection. In H.Õim (ed), Estonian in the Changing World. University of Tartu. Department of General Linguistics. Tartu, 1996, pp. 7-32.

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

Hennoste, Tiit; Muischnek, Kadri 2000. Eesti kirjakeele korpuse tekstivaliku ja märgendamise printsiibid ja kahe allkeele võrdlemise katse. Arvutuslingvistikalt inimesele. Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1. Toim. T.Hennoste. Tartu 2000, lk 245-284.

Kaalep, Heiki-Jaan (1997) An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. Computers and the Humanities 31, pp. 115—133.

J. Sinclair. Corpus Conordance Collocation. Describing English Language. Oxford University Press 1991.