The Estonian Reference Corpus: its composition and morphology-aware user interface

Heiki-Jaan KAALEP[1] , Kadri MUISCHNEK, Kristel UIBOAED and Kaarel VESKIS
University of Tartu

Abstract. This paper gives a brief overview of the composition as well as technical and morphological annotation of the Reference Corpus of Estonian. A user interface using the morphological information about lemmas and grammatical categories of word-forms is presented.

Keywords. corpus compilation and mark-up, corpus user interface, morphological annotation

Introduction

The Estonian Reference Corpus[2] is a collection of written present-day language consisting of ca 245 million words at the moment. In our paper we are going to describe the overall composition of the corpus; say a few words about its technical and morphological annotation and present a corpus query system based on morphologically analyzed version of the corpus. [3]

1. The overall design and technical annotation of the Corpus

The Estonian Reference Corpus is a non-balanced one: the newspaper texts make up 75% of the Corpus, fiction texts 2%, scientific texts 2%, legalese 5%, parliament transcripts 5% and the texts of the "new media" 9% of the corpus. By "new media" we mean the genres of the computer-mediated discourse; i.e. the chatrooms (Internet relay chats), Internet forums, newsgroups and user comments from the news portals.

The technical annotation of the Corpus follows the TEI guidelines. The traditional written texts (i.e. newspapers, fiction texts etc) are annotated for the text structure. Non-textual material (graphs, formulae, pictures, tables etc) has been omitted and represented by a tag <gap desc='description_of_the_omitted_material'>.

[1] Corresponding Author.

[2] http://www.cl.ut.ee/korpused/segakorpus/index.php?lang=en

[3] This work was supported by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS and by the Estonian Ministry of Education and Science (grant SF0180078s08)

The annotation of the "new media" texts is different from that of the rest of the Corpus. The basic idea behind tagging was that the transcript of a chatroom or a newsgroup or the text of an online forum is similar to a transcript of a drama play: the actors enter the stage, produce their lines, and leave the stage. Thus, the time of the text entered to a web site, if available, has been tagged as <time>, the speaker as <speaker>, a text of one speaker as <sp>, the theme or title of the message as <head>, and the actions between the chat lines as <stage>.

The mark-up of the Corpus follows the currently outdated P3 version of the TEI guidelines[4] that has some significant disadvantages compared to later versions of TEI. In 2002, TEI changed its underlying representation from SGML to XML with the P4 version and in 2007; the P5 version added some architectural changes. There are a number of benefits in switching from SGML to XML, one of which is that XML has a number of standards and specifications that SGML lacks.

Our plan is to migrate from P3 to P5, not skipping the P4 stage, but instead use P4 as an intermediary stage in order to facilitate the migration. The reason is that in 2002, The TEI Task Force on SGML to XML Migration has devised a Practical Guide to Migration of TEI Documents from P3 to P4[5]. TEI has also guidelines for migrating from P4 to P5[6] but no guidelines for direct P3 to P5 migration are known to us. As a part of the migration process, we have already converted the text encoding of the corpus from ASCII and SGML entities to UTF-8.

2. Morphological annotation of the Corpus

Estonian is an agglutinating language; it uses inflection for encoding the syntactic relations between the words of a sentence. At the same time Estonian has some fusional traits: it has a tendency to fuse morphemes so that they are difficult to segment. For example the first four case-forms of the word *käsi* 'hand' would be in singular *käsi käe kätt kätte* and in plural *käed käte käsi kätesse*. That entails the necessity of morphological analysis for a corpus query system, as in many cases it is not possible to retrieve all inflectional forms of a word using its base form and some kind of regular expression. To make matters worse, 45% of tokens in a text corpus can be analysed in several ways, if the context they occur in is not taken into account. In other words, 45% of the tokens are morphologically ambiguous.

The corpus has been annotated morphologically by Filosoft Ltd. using their morphological analyzer (including guesser for out-of-dictionary words) of Estonian and a HMM disambiguator. The principles of the approach date back to [1], but the tools have been developed further, e.g. the HMM disambiguator has been implemented as a trigram HMM and trained on a manually annotated corpus of 500,000 tokens[7] . The categories used by the morphological analyzer and disambiguator are based on [2] .

After disambiguation, 10% of the tokens still remain ambiguous. This is because if we do not have rules or data for choosing the right annotation with a high probability, it is better not to make the choice at all. The ambiguous tokens fall into the following categories: participles (ambiguous between verb and adjective readings), pronouns

[4] https://docs.google.com/viewer?url=http://www.tei-c.org/Vault/GL/p4beta.pdf

[5] http://www.tei-c.org/Activities/Workgroups/MI/index.xml

[6] http://www.tei-c.org/Guidelines/P5/migrate.xml

[7] http://www.cl.ut.ee/korpused/morfkorpus/

(ambiguous between singular and plural, and between pronoun and numeral readings), verb form *on* (ambiguous between '(he) is' and '(they) are'), uninflected words like *kui*, *otsekui*, *nagu* 'if', 'as if', *just* 'just' (ambiguous between conjunction, interjection and adverb readings).

An evaluation of the quality of the disambiguation showed that depending on the text class, 3-6% of the annotations were not quite correct[8]. An error could be in the lemma form, inflectional category, or word class. The evaluation also showed that if the HMM disambiguator was used on a text class, not seen in the training phase, the quality of its output was about 1 percent point lower. Thus we expect the quality to remain rather stable across the whole annotated corpus.

The "new media" subcorpus has not been morphologically annotated yet, the reason being that the texts of the new media, especially those of the chatrooms, contain a lot of word-forms not occurring in the texts of the standard written language or being used in different function and meaning. Due to these facts these texts need some special pre-processing prior to the morphological analysis and the lexicon of the morphological analyzer needs to be customised.

3. The morphology-aware user interface

For years, the whole corpus has been freely downloadable for non-commercial purposes, with the possibility to use software of any origin for doing research on the downloaded texts. However, feedback from potential users has indicated that Estonian linguists would prefer to query the corpus via Internet, using a simple search facility instead. So it is necessary to have an appropriate corpus toolkit, commonly known as concordance software, which enables the user to query the corpus. For several years our corpus has had a simple search facility[9] that retrieves a sequence of symbols from the corpus. Another, new interface[10] enables the users to query the corpus using the lemmas of word-forms and/or morphological information, in combination with surface word forms.

The new search facility is trying to be balanced between 1) the ease of use, 2) the functionality, required by linguists, and 3) simplicity of the maintenance (including upgrading) of the corpus and the software.

The internal representation of the Corpus for ensuring fast retrieval was designed and implemented by Rene Prillop, who also designed and implemented the query interface. The conversion from TEI format to the morphologically annotated one was performed by Tarmo Vaino.

Figure 1 shows the first 5 results for a query for a multi-word verbal expression *silmi lahti hoidma* lit. 'keep one's eyes open', i.e. 'pay attention, be watchful'. The query was submitted to several sub-corpora (shown as tabs on the web page) at the same time, but only the results from one subcorpus are shown – 119 sentences from the newspaper *Eesti Päevaleht*. For every sentence, there is a clickable field for showing the exact source of this sentence. The searched terms are highlighted.

[8] http://teataja.ee/veskis-liba-syntax-assignment-modified.pdf

[9] http://www.cl.ut.ee/korpused/kasutajaliides/

[10] http://www.keeleveeb.ee

Figure 1. Results for a query *silmi lahti hoidma.*

Note that the order of the searched word forms may vary and that there may be intervening words. The second resulting sentence shows what happens when one clicks on a word (e.g. *silmad 'eyes'*): its morphological analysis – lemma (*silm*), inflectional ending (*d*), word class (*S* – noun) and grammatical information (plural nominative) – are displayed.

The query can be submitted via a set of text fields (for word forms, lemmas, parts of words) and clickable boxes. The user input is transformed into a query string (*silm@l lahti hoidma@l* on Figure 1), which is then used by the system to perform the search. The user can save this string, so that next time she need not tick the same boxes again, but may paste the saved string directly to the search box.

4. Conclusion

This paper gave a short overview of the Reference Corpus of Estonian, its annotation and a new morphology-aware user interface. Our plans for the near future are twofold: first, to perform the morphological annotation of the "new media" texts and thus make them usable via a search interface. Second, we are working on splitting the sentences into clauses in order to provide better context for retrieving co-occurrences of words.

References

[1] Heiki-Jaan Kaalep, Tarmo Vaino. 2001 Complete Morphological Analysis in the Linguist's Toolbox. Congressus Nonus Internationalis Fenno-Ugristarum Pars V, pp. 9-16, Tartu.
[2] Viks, Ü. 1992. Väike vormisõnastik. ETA EKI, Tallinn