Process Model for Composing High-quality Text Corpora

Mikko Lounela

Research Institute for the Languages of Finland
Sörnäisten Rantatie 25, FI-00500, Helsinki, Finland
mikko.lounela@kotus.fi

Abstract

The Teko corpus composing model offers a decentralized, dynamic way of collecting high-quality text corpora for linguistic research. The resulting corpus consists of independent text sets. The sets are composed in cooperation with linguistic research projects, so each of them responds to a specific research need. The corpora are morphologically annotated and XML-based, with in-built compatibilty with the Kaino user interface used in the corpus server of the Research Institute for the Languages of Finland. Furthermore, software for extracting standard quantitative reports from the text sets has been created during the project. The paper describes the project, and estimates its benefits and problems. It also gives an overview of the technical qualities of the corpora and corpus interface connected to the Teko project.

1. Introduction

Text corpora are usually composed in a concentrated manner; the sampling principles, collection methods, markup, and annotation principles and techniques are designed, after which the texts are collected and processed to form a consistent, representative corpus. The corpora may be augmented afterwards, usually according to the original sampling and design principles. In this process model, the principles of choosing the texts, and the manner and extent of markup and annotation are entirely in the hands of the institution collecting the corpus. The users will have to adjust to the composers' decisions. This model has an effect on the way the researchers use the corpus: how and what they research, and to what extent they can use corpora in their research if they do not adjust their research questions to available corpora.

2. Goal of the paper

This paper presents an alternative, decentralized process model for composing a research text corpus. In this model the primary end users of the corpus can direct the collection of the texts and influence the markup so that the corpus will adapt to their specific needs as well as possible. This requires the involvement of linguist researchers in the process of composing the corpus, which may not be easy for a few reasons. In the light of the usability of the corpus in linguistic projects, it is, however, worth it.

3. Project

The alternative process has been tested in a pilot project for collecting a corpus of Modern Finnish in the Research Institute for the Languages of Finland. The project was called Teko Project, so I am calling the model Teko Model, and the corpora collected according to it Teko corpora. Finnish is a small language, for which there are no parsers of sufficient quality for scientific research. This has an obvious effect on the process of composing an annotated corpus of this type.

In the pilot project the division of labour between linguists and the corpus people has been roughly the following: the linguists in the research project choose and collect (and possibly digitize if necessary) the texts. Estimating the size for the sub-corpus is done in cooperation with the corpus people. The corpus people convert the texts to sentence-level TEI-P4 XML. The sentence structure is revised by the linguists, and the optional research project specific markup is added cooperatively. This may include marking up items such as names, numbers, and special text sequences. The revised XML document is run through a morphological analyzer by the corpus people, and the result is disambiguated and augmented manually by the linguists. In a related project, even word dependencies were added by hand to a corpus of Early Finnish.

Finally, the corpus people add meta-data to the text collection and attach them to the main corpus. When the corpus is ready, it is run through a set of linguistic report generators which create quantitative data (consisting mainly of frequency lists and other distributional figures) for selected linguistic features occurring in the texts (Lounela, 2005). As this process is repeated in cooperation with different linguistic research projects, a considerable body of carefully selected corpus texts of high quality can be collected for the benefit of the research community (along with the comparable quantitative data describing each of them). The method has been used as part of a number of research projects in its different developmental phases during the past years (Heikkinen, 1999; Kankaanpää, 2006; Tiililä, 2007).

4. Corpus

4.1. Design

A Teko corpus is inherently modular, and dynamic in the sense that it grows naturally. Its design is strict, and yet flexible on the other. The main corpus consists of sub-corpora of moderate size. The sampling principles along with the standards and techniques used in markup and annotation are clearly defined. Additional markup and annotation may be added in accordance with the standards, if the linguistic research project needs that. In the following, some of the design principles will be considered:

- Choosing the texts: Each sub-corpus consists of a set

of texts that are relevant to the research project. The size of the sub-corpus is largely determined by the resources of the project.

- Sampling: The corpus consists of whole texts, so no information about the text flow is lost.
- Encoding: Markup is performed according to (slightly moderated) TEI-P4 XML definition. There is a required minimum level for the text structure markup (word level). Any TEI-P4 conformant markup can be added on top of that.
- Annotation: All the texts in the corpus are run through a morphological analyzer (TWOL, 2007), and disambiguated and augmented by hand. This phase is time-consuming and can be done differently on corpora of language with high-quality parsers. The markup and annotation are explained in an earlier publication (Lehtinen and Lounela, 2004). The example below gives a view of the result of text annotation (Roughly translated *This time I am speaking about turism*.)

  <w id="w484" lemma="tm" norm="tällä" type="PRON" msd=" DEM ADE SG ">Tällä</w>
  <w id="w485" lemma="kerta" norm="kertaa" type="N" msd=" PTV SG ">kertaa</w>
  <w id="w486" lemma="puhua" norm="puhun" type="V" msd=" PRES ACT SG1 ">puhun</w>
  <w id="w487" lemma="matkailu" norm="matkailusta" type="N" msd=" DV-ILE DV-U ELA SG ">matkailusta</w>
  <w id="w488" lemma="." norm="." type="PUNCT" msd=" FULLSTOP ">.</w>

- Meta-data: Standard meta-data are attached to each text document and sub-corpus at different levels of the corpus tree. These meta-data work as the backbone of the corpus; they compile unrelated text sets to a consistent, modular text corpus. Figure 1 gives an overview of the metadata system. For sub-corpora, the meta-data differ somewhat, eg., instead of a link to the digitized image of the original text, there is al link to the texts, and instead of a link to the text set meta-data, there is a link to supercorpus meta-data.

The resulting corpus is a collection of text sets, each representing itself. In addition to the morphologically annotated texts, the corpus includes basic meta-information using the DCMI meta-data recommendation, expressed in the RDF/XML format (Beckett, 2002) with some modifications. The modifications are made mainly in order to achieve applicability as a browsing environment (Lounela, 2007). As each text set is part of a research project, the resulting publications of the project can also be linked to the meta-data, possibly in connection with other relevant publications, which makes the corpus even more usable for researchers (and the results more verifiable than those of linguistic research in general). The generated quantitative linguistic reports can also be linked to the text sets through the meta-data system. The resulting whole consists totally of organized XML or HTML files and style sheets, and it can be served to the general public through the www.

4.2. Sampling and Representativeness

The Teko model has its own principles of both sampling and representativeness. Both of these are based on the view that the less the composers of the corpus make hypotheses about language and text, the more freely these hypotheses can be made by researchers. All the ways of sampling the texts, or balancing the corpus, involve such hypotheses (or generalizations), and they invite the researchers to base their research on the same assumptions than the corpus builders have. Such assumptions may include views about what types of texts represent "real" language or that the structures of individual texts are not important in linguistic research.

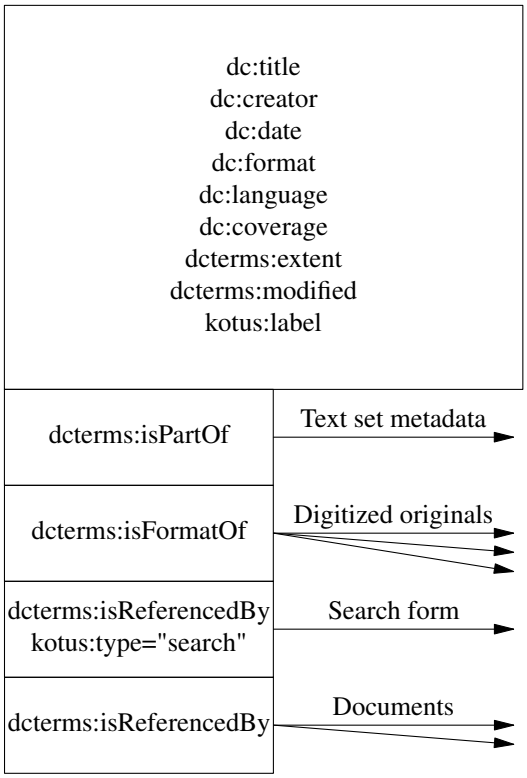

Figure 1: Corpus metadata

The Teko model does not apply any techniques for sampling the texts. All the texts are included in the corpora as they are. Naturally, digitizing and normalizing is necessary, and furthermore, adding the XML structure and morphological analysis includes its own preassumptions of text and language (in this case, especially the Finnish language). These assumptions can, however, be challenged, because the whole text is included in the corpus. To make it possible to let the researcher decide (as far as possible) which properties of the texts are relevant, we also include digitized images (or other originals, eg., original www-pages) of the texts to the corpus, if possible.

The model does not select the texts to be included in the corpus. Our view of representativeness is that no prechosen text set can represent a language or its use. A Teko corpus is an organically growing collection of free-standing text sets. These text sets do not represent any genre, text type or other such abstract concept. They represent them-

selves, and can be examined as they are. This view is the opposite of most other approaches to corpus composition, eg., in DiaCORIS and BNC (Onelli et al., 2006; Ashton, 1996).
The selection of the text and text sets is done by linguist researchers who want to do research on them. However, some basic sets will be more likely to be used for comparing the results of new sets trhan others. In the current corpus, these include the newspaper news text set (and, in the future probably, the diachronic weekly journal text set). Table 1 shows the current structure of the corpus. The major part of the corpus consists of restricted text sets. These sets are subject to copyright, so we are not allowed to republish them openly on the web.

4.3. User Interface

The Teko model is compatible with the Kaino text corpus user interface. The diachronic Kaino corpora include 11 600 000 words of written text from the sixteenth century to present day. The meta-data of each subcorpus (and subsubcorpus) and text file is automatically transformed into a web page[1]. As the meta-data of the corpus parts refer to each other, they build a browsable tree structure of different corpora (Lounela, 2007).
The corpus interface also includes a search engine. Each sub-corpus of the Kaino corpus server has its own search page, which makes it possible to search words and word combinations using regular expressions or simple wildcard characters. Also, the search form enables choosing a text set, the form of the search result (examples, concordance, or frequency list), choosing the maximum distance of the search words, and making a choice between showing all the matches or a random subset of them[2].

4.4. Current state

At the moment the Teko corpus includes 485 965 words in 8 text sets. One major text set (680 936 words) is under preparation. We are also considerin to compose another major text set consisting of legislative texts.
The main tasks before finishing the pilot project are

- publishing the restricted text sets through the Kaino search form system.
- preparing the meta-data trees of all (even restricted) text sets and publishing them through the Kaino browsing system, and
- preparing and publishing the quantitative morphological reports about all text sets through the Kaino browsing system.

5. Problems

The main problems that were found during the pilot project involved copyright and data security issues (regarding the publicizing of the texts); motivating the linguists to participate in the project; and managing the quality control of the work phases.
The first problem is an issue that touches the field of corpus linguistics very widely, and it cannot be solved by the researchers. There are, however, certain actions that can be taken. The copyright problems can be tackled by choosing copyright-free texts, if possible, and by using model contracts that can be used to obtain the right for (possibly limited) republishing of the texts. If republishing is impossible, the research results can be published along with the quantitative data, and the texts can be shared with interested researchers privately, without republishing them. The pilot project collected two copyright free text set - a collection of New Year's speeches by the presidents of Finland and a collection of legislative texts. These collections are published online in Kaino text corpus system (KAINO, 2007). Other materials (see table 1 for all the text sets) are not published due to copyright restrictions. They may be made available legally through the search engine which delivers them in a way that is clearly not a violation against the copyright laws. In the legal sense, this should be equal to the old technique of sampling the texts before publishing the corpus. At the moment, other than the two published text sets are not included in the range of texts available through the search engine, but this will change in the future.
The main data security problem in our environment concerns our policy regarding possible editing and republishing of the texts by another party. As the published texts are not bound by the copyright law, this issue is not of primary concern, even though we have the copyright to the markup and design of the corpus. At the moment, the texts and the meta-data are published without license, but some licensing system such as GPL (Smith, 2007) is under consideration.
The second problem concerns motivating the linguists. Linguist will hardly have this type of an effort in mind while building up a humanistic research project, and the work required to have a reasonable text set disambiguated by hand is considerable. However, the process offer opportunities of

- having a design corpus,
- obtaining a secure, standard set of quantitative data,
- adding special markup to the texts and having quantitative data for entities of special interest (eg. names, numbers, special text sequences), and
- having comparable figures from several other text sets concerning the features presented in the quantitative standard reports.

These options should be quite attractive once the linguists understand them. In addition, the markup and disambiguation work itself offers a new point of view to the research material, and can help revising the research questions. The problems seem to be mostly related to informing the researchers at the right phase of the project and overcoming the suspicion about new ways of working.
The quality control problems in the pilot project occurred mostly because we used changing workforce in the markup

[1] For an example of rendered metadata, see http://kaino.kotus.fi/korpus/meta/korpus_coll_rdf.xml
[2] For an example of a search form, see http://kaino.kotus.fi/korpushaku/teko-haku.xql.

Text set	Size (words)	Time span	State
Texts from a weekly journal	680 936	1917 – 1972	Restricted (and under preparation)
Presidents' New Year's speeches	63 110	1935 – 2007	Free
Administrative press releases	19 065	1979 – 1999	Restricted
News on plain language	14 530	2001 – 2003	Restricted
Guidelines given by church administration	17 639	2002	Restricted
Short news from local newspapers	97 325	2002	Restricted
Handbooks by tax administration	18 591	2002	Restricted
Laws and directives	232 449	2002 – 2003	Free
Communal introductory www-pages	23 256	2004	Restricted

Table 1: The text sets

and disambiguation, and because of the pilot nature of the project. The first problem can be overcome by having well-designed, strict and detailed guidelines concerning text structure and borderline cases in the disambiguation work. The second problem can be tackled by allocating enough resources to guidance, and connecting experienced disambiguators and structurers to new projects.

6. Discussion

The presented Teko model presents a decentralized way of collecting annotated high-quality corpora for the individual linguistic projects. The size of the text material and hence the required effort can be scaled according to the size of the project.

The Teko model has some benefits as compared to the dominating large, centralized models of corpus composition. These include:

- The collected corpus meets the special needs of the research project.
- The model includes a standard procedure for extracting quantitative data from individual text sets. The data can be easily compared with data from other text sets.
- The model offers a way of composing corpora with few presuppositions about language (coverage), text (normalizing), or linguistics.
- The model offers a way of composing corpora with a moderate amount of persistent effort as an alternative to big, short-lived projects.

However, the model is best suited to projects that do not intend to involve generalisations about language (human language or, eg., the Finnish language), or other such linguistic abstractions as their results. The most obvious user groups have thus far consisted of researchers doing (critical) text analysis, but there are many reasons to believe that the model can benefit all linguistic research.

7. References

Guy Ashton. 1996. *The British National Corpus as a language learner resource*. In S. Botley, J. Glass, T. McEnery and A. Wilson (editors): *Proceedings of Teaching and Language Corpora 1996*, pages 178 – 191, Lancaster, UCREL.

Dave Beckett. 2002. *Expressing Simple Dublin Core in RDF/XML*. Dublin core Metadata Initiative. Online: http://dublincore.org/documents/dcmes-xml/. Referred 1.10.2007.

Vesa Heikkinen. 1999. *Ideologinen merkitys kriittisen tekstintutkimuksen teoriassa ja käytännössä* [Ideological meaning in the theory and practice of critical text analysis] (In Finnish). Helsinki, Suomalaisen Kirjallisuuden Seura.

Salli kankaanpää. 2006. *Hallinnon lehdistötiedotteiden kieli* [Language of Administrative Press Releases] (In Finnish). Helsinki, Suomalaisen Kirjallisuuden Seura.

Outi Lehtinen and Mikko Lounela. 2004. A model for composing and (re-)using text materials for linguistic research. In Marja Nenonen, editor, *Papers from the 30th Finnish Conference of Linguistics*, pages 73 – 78, Joensuu, University of Joensuu.

Lingsoft, Inc. 2007. *FINTWOL: Suomen morfologinen jäsennin* [FINTWOL: Morphological Parser for Finnish] Online: http://www2.lingsoft.fi/cgi-bin/fintwol. Referred 1.10.2007. (In Finnish).

Mikko Lounela. 2005. Exploring morphologically analysed text material. In Antti Arppe etc. (editors): *Inquiries into words, constraints and contexts. Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*, pages 359 – 267, Helsinki, Gummerus,

Mikko Lounela. 2007. Anatomy of an XML-based text corpus server. In Joakim nivre etc., editors, *Nodalida 2007 Proceedings*, Tartu, University of Tartu, CD-ROM.

Research Institute for the Languages of Finland. 2007. *Kotuksen tekstikorpuksia* [RILF Text Corpora]. Online: http://kaino.kotus.fi/korpus/meta/korpus_coll_rdf.xml. Referred 15.3.2007. (In Finnish).

Brett Smith. 2007. *A Quick Guide to GPLv3*. Free Software Foundation, Inc. Online: http://www.gnu.org/licenses/quick-guide-gplv3.html. Referred 4.3.2008.

C. Onelli, D. Proietti, C. Seidenari, and F. Tamburini. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of LREC-2006, The Fifth International Conference on Language Resources and Evaluation*, Genoa, pages 1212 – 1215.

Ulla Tiililä. 2007. *Tekstit viraston työssä* [Texts in a bureau's work] (In Finnish). Helsinki, Suomalaisen Kirjallisuuden Seura.