Anatomy of an XML-based Text Corpus Server

Mikko Lounela
Research Institute for the Languages of Finland
Sörnäisten rantatie 25
00500 Helsinki
Finland
mikko.lounela@kotus.fi

Abstract

This document describes an XML-based data model for annotated, modular text corpora along with a WWW-interface for browsing such corpora, reading the texts, searching for examples, and extracting information of word usages. The interface is based solely on programs and techniques belonging to the XML-family. The corpus model is designed in such a way that new parts (texts, sub-corpora) can be easily plugged in the system as far as they fit into the model. Furthermore, the model includes slots for such parts that are not conventionally included in text corpora. These may include (digitized) originals of the texts and links to other relevant documents. The searching interface for the model is based on XML query language that enables the developers to add queries to the system for extracting detailed linguistic information from the texts, depending on their annotation level. The corpus model and its interface can be seen as a step towards a general quantitative tool for text linguistic research, including the data model and programs for browsing, querying, and analyzing the texts.

1 Credits

I am indebted to Outi Lehtinen, who has worked on developing the database environment for the searching interface, and helped in numerous other ways. I am also indebted to Mikko Virtanen, who has carried out a lot of important work in encoding the corpus and developing the style sheets for the corpus interface. He has also designed modifications for the metadata model.

2 Introduction

The Research Institute for the Languages of Finland (RILF) has developed a text corpus data model in recent years. The model is based on de facto standards belonging to the XML family, which offers a well-defined, consistent set of languages for defining and processing structured documents. Further, the processes modelled by XML-related definitions can often be executed using free, high quality applications. The main goal of this work has been to develop a consistent, easy-to-use model for text collections, text documents, and their metadata. The corpus model is easy for the developers in the sense that new texts and sub-corpora can be added to the collections with little effort. Browsing and searching applications are also relatively easy to build and maintain. As a matter of fact, the browsing interface is implemented by adding style sheets to the metadata and text files. The searching interface can be implemented in many ways. The current searching interface is built on top of an XML database. The ease for developers should naturally lead to ease of use for linguist researchers and other corpus end users.

The model has been implemented as part of a web-based text corpus service, belonging to the RILF data service Kaino (RILF, visited 27.2.2007). The service was opened to public in December

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 337–344

2006[1]. The text corpus interface of Kaino service is divided into two parts: a browsing interface, which is built on top of the file system of the corpus server, and a searching interface, which uses a free XML database as the search engine. The interfaces are cross-linked. The definitions used in the system are well-known and XML-based, being very close to de facto standards. The tools used in the system are free and have an open source code.

In this article, I will present the data models and techniques included in the corpus model, as well as the implementation of the model. First, I will present the XML structure of the corpus texts, and the metadata model of the corpus. Then I will go through the technicalities of the current browsing interface and searching interface. After that I will introduce briefly the corpora included in the RILF text corpus collection, and their annotation and special features. Finally, I will consider the expandability of the interface and some future prospects for the model.

The documents referred to in this paper include web pages of institutions, in which case the authors are anonymous, and the publishing dates unrevealed. I refer to these documents using the institutions' names instead of authors' names, and visiting dates instead of publishing years. Some of the references concern research carried out in Finnish language, in which case the publications are in Finnish.

3 Corpus Data Structure

Figure 1 presents an overview of the modular corpus model. Metadata files include descriptions of the corpus, its sub-corpora, and individual texts belonging to the sub-corpora, as well as links that connect the parts to each other. New parts can be added to the corpus simply by inserting links pointing to them into the description file at suitable level of the corpus structure. At the moment, the RILF text corpus collection includes generally four levels. There can be more levels of collections in the corpus. In principle the corpus tree can be of any depth.

The links connect the corpora to their parts, each being represented by their metadata files. In the top of the tree, the metadata descriptions of the text files are linked to the corpus text files. In some text collections, the descriptions are also linked to the digitized originals of the texts. Links to any other versions of the texts can be added to the descriptions, as well as links to any other relevant documents.

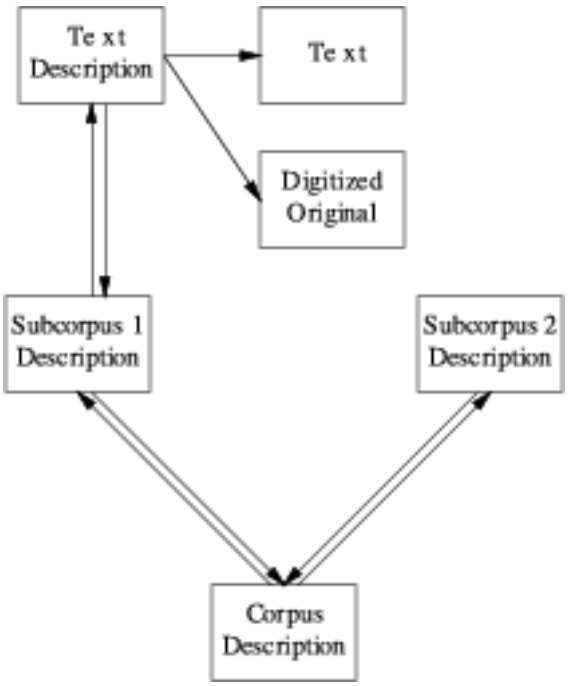

Figure 1: Overall corpus structure

3.1 Text Structure

The text structure of the corpus model is based on Text Encoding Initiative TEI P4 XML data model (TEI, visited 27.2.2007 a). Depending on the subcorpus, the texts may be encoded to the paragraph level, sentence level or word level. The word level analysis includes morpho-syntactic tagging, and possibly semantic elements. The semantic elements may denote persons, names, dates, addresses etc. In our experience, almost any element may need further classification, when texts are used for text linguistic research. The needs above require some alternations to the original TEI structure (eg. an attribute for the word element to carry the morpho-syntactic description, and a global type attribute for sub-classifying any element). On modifications to

[1]The text corpus interface is can be found at http://kaino.kotus.fi/tekstikorpukset/.

the TEI P4 structure for research corpora, see Lehtinen and Lounela (2004).

Some sub-corpora include texts with alternative structures, such as dictionary, drama and poetry. For these, customized structure definitions (DTDs) have been created using the automatic DTD generator TEI Pizza Chef (TEI, visited 27.2.2007 b). These DTDs have not been modified.

3.2 Metadata Structure

The metadata model of the RILF text corpus collection uses the Dublin Core metadata Initiative (DCMI, visited 27.2.2007) element set, expressed in the Resource Description Framework (W3C, visited 27.2.2007 a) data model. About the DCMI metadata structure, expressed in the RDF/XML format, see Beckett et al., (2002) and Kokkelink and Schwänzl, (2002). A model for using these techniques to form a metadata backbone for a structured corpus has been outlined in Lounela (2002).

We have made some additions to the information content of the DCMI element set. The motivation for them is to make it more convenient to browse the corpus. The additions consist of including one extra element and some attributes into the definition. The element included in the descriptions is kotus:label, and it is used to label a (sub)corpus, a text, or a link in the browsing interface. Links are presented as dc:Relation-elements. In the browsing system, a relation-element contains two elements: a selected DCMI-element that expresses the URL and classifies the relation, and a kotus:label-element for labeling the link. The DCMI elements classifying the relations used in the RILF text corpus collection are the following.

- dcterms:isPartOf for linking a collection to its super-collection.
- dcterms:hasPart for linking a collection to its sub-collection.
- dcterms:isReferencedBy for linking a collection or a text either to a search engine or an external document of some relevance.
- dcterms:isFormatOf for linking a text document to its digitized original.

The attributes added to the DCMI/RDF/XML for use in the RILF corpus model add some information to the existing elements. This information typically helps in grouping some links or adds linking information to the browsing system. The added attributes are the following.

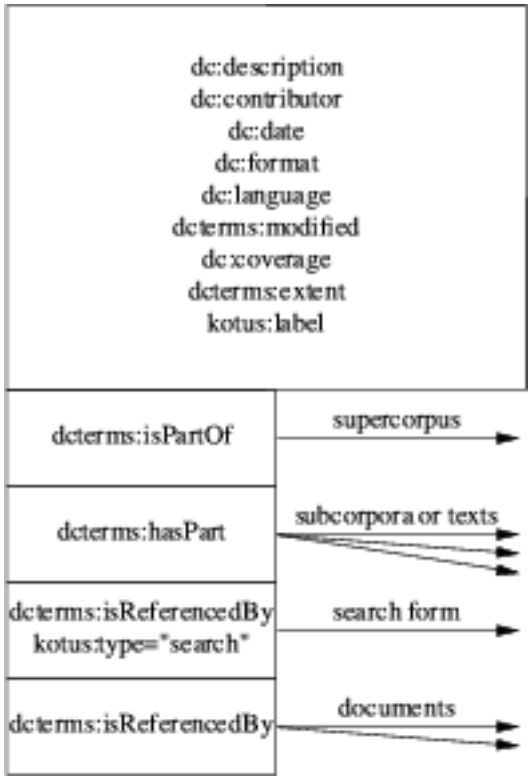

Figure 2: Meta-information of a sub-corpus

- kotus:unit-attribute to qualify the dcterms:extent-element. In corpora, the unit is normally a word.
- kotus:type-attribute is added to the DCMI elements classifying the relations. At the moment, type-attribute is used for separating links to the searching interface from links to other documents, as all of these are expressed by dcterms:isReferencedBy-element.
- kotus:metalink-attribute for adding an URL of a metadata description of a resource to the relation-classifying elements.

- kotus:class-attribute to group the sub-corpora of a corpus and label the groups. The sub-corpora are referenced using dcterms:hasPart-elements.

Figure 2 illustrates the meta-information structure of a sub-corpus in the RILF corpus system. The large box in the figure includes the elements used to describe the current collection. The small boxes and arrows below the large box represent linking that implements the hierarchical corpus structure. Links to other relevant documents (expressed by dcterms:isReferencedBy-element) are optional. Description of a text document has no links to sub-corpora, but it may have dcterms:isFormatOf-links to digitized originals. A text document has a dc:title-element that expresses the original title of the text.

The following code extract represents a meta description of a text document. After the XML formalities, it contains an rdf:RDF-element that includes all the DCMI elements describing the current document, and the links forming the browsing system. The links are expressed as dc:Relation-elements. The first of them links the text to the search form (teko-haku.xql), the second one links the text to the collection in which it belongs to (Speeches of President Halonen), and the third one links it to its original HTML version on the web.

The text content of the elements and links in the example are edited for presentation purposes. The original meta description can be found in the Kaino data service.

```
<?xml version="1.0"
 encoding="iso-8859-1" standalone="yes"?>
<?xml-stylesheet
 type="text/xsl" href="/tyyli/dc_teksti.xsl"?>

<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:kotus="http://www.kotus.fi/"
 xmlns:dcterms="http://purl.org/dc/terms/">

<rdf:Description
 rdf:about="../halonen/halonen_2003.xml">

<dc:title>New Year's Speech 2003</dc:title>
<dc:creator>Halonen, Tarja</dc:creator>
<dc:date xml:lang="FI">2006</dc:date>
<dc:format>TEXT/XML</dc:format>
<dc:language>FI</dc:language>
<dc:coverage>1.1.2003</dc:coverage>
<dcterms:extent
 kotus:unit="words">985</dcterms:extent>
<dcterms:modified>20.12.2006</dcterms:modified>
<kotus:label>New Year's Speech 2003</kotus:label>

<dc:Relation>
<rdf:Description>
<dcterms:isReferencedBy
 kotus:type="search"
 rdf:resource="/korpushaku/teko-haku.xql"/>
<kotus:label>Search Form</kotus:label>
</rdf:Description>
</dc:Relation>

<dc:Relation>
<rdf:Description>
<dcterms:isPartOf
 rdf:resource="../teksti/presidentti/halonen/"
 kotus:metalink="../halonen_coll_rdf.xml"/>
<kotus:label>Speeches of President Halonen</kotus:label>
</rdf:Description>
</dc:Relation>

<dc:Relation>
<rdf:Description>
<dcterms:isFormatOf kotus:type="external_link"
 rdf:resource="http://www.tpk.fi/"/>
<kotus:label>HTML text</kotus:label>
</rdf:Description>
</dc:Relation>

</rdf:Description>
</rdf:RDF>
```

4 Browsing Interface

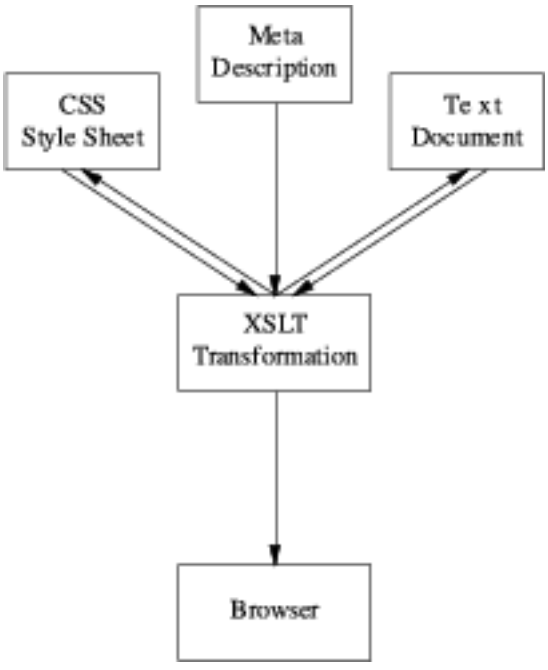

Figure 3: Browsing interface

The browsing interface for the corpus model described above is implemented in XSLT transformation language (W3C, visited 27.2.2007 b) and

Cascading Style Sheet language CSS (W3C, visited 27.2.2007 c). Figure 3 illustrates the browsing system, in which the meta descriptions are attached to XSLT transformation definitions that rearrange them and transform them to HTML documents. These documents get their appearance from CSS style sheets. In the case of text documents, XSLT transformations rearrange meta descriptions and fetch corresponding text documents to be displayed along with them.

This arrangement is modular in the sense that new material can be easily plugged in at any level of the system. Because the transformations are performed dynamically by the web browser, the corpus can be browsed on the spot, as it is, with all its meta information and text content

5 Searching Interface

The searching interface of the current RILF text corpus collection is built on top of eXist XML database (Meier, 2007). The queries are implemented in Xquery language, using some database-specific extensions. When a query is carried out, the text documents of the sub-corpus to be queried are identified and located using the meta descriptions of the documents. After that, the documents are queried one by one.

Figure 4 illustrates the basic functions of the searching interface. Search form sends a query to the database, where the search is performed. The resulting elements are formatted with an XSLT transformation and embedded in the reloaded query form. The XSLT transformation links search results to the corresponding passages of text in the browsing system.

Using Xquery makes it straightforward to use meta information included in the DCMI/RDF descriptions for filtering text documents to be queried. This makes it possible to use publishing date or language of a text as a key for fetching documents. At present, in the RILF text corpora, the publishing dates of the old texts are not consistently accurate, and all the texts are in Finnish, so this feature is not implemented in the current version of the system.

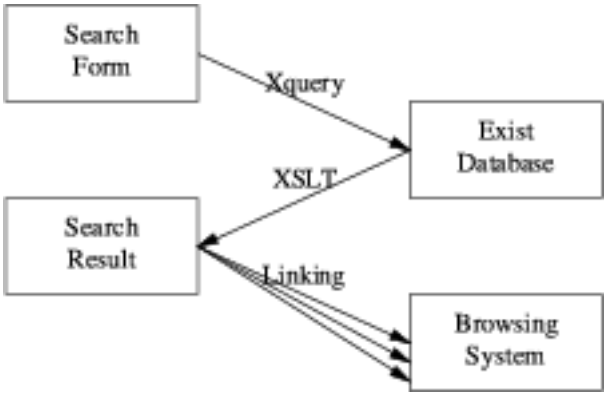

Figure 4: Searching interface

Xquery, being close to a full programming language, offers good options to process the query results further. At the moment the searching interface offers a possibility for producing a frequency list of occurrences of the queried word in a sub-corpus, but there are many more ways of producing summaries of linguistic features of the sub-corpora (or sets of selected text, if that feature will be implemented).

6 RILF Text Corpus Collection

The system, including separate but interlinked browsing and searching interfaces, is currently implemented at RILF. Publicly available parts of the text corpora possessed by RILF are served through it. The major part of the RILF collections form a diachronic corpus collection of Finnish literary texts. The oldest collected texts date back to the birth of the Old Finnish Literary Language in the sixteenth century. The collection goes diachronically through the Early Modern Finnish Literary Language from the nineteenth century to the present day Literary Finnish. Because of copyright issues, the public texts of the modern language available in the Kaino data service are very few. In addition to the diachronic, literary texts, the corpus includes a collection of Finnish proverbs, collected in the 1930s.

The current structure of the public RILF text corpus collection is presented in the list below (the names of the corpora on the list are not official). The corpus collection consists of different levels as explained above. The corpus collection root contains the diachronic corpora and the proverb collection. These corpora are further divided into sub-

collections. The nature of the division depends on the corpus. In the historical corpora of old texts (corpora of Old Literary Finnish and Early Modern Finnish and the collection of Literary Classics), the division into sub-collections is mainly based on authorship. In the proverb collection, the division is regional. In the collection of texts in Modern Finnish, the division is based on text types, and the sub-corpora may be divided further according to author, source or some other property.

- The RILF Text Corpus Collection
  - The Corpus of Old Literary Finnish
    - 12 sub-collections
    - 132 texts
    - 3.5 million words
  - The Corpus of Early Modern Finnish
    - 94 sub-collections
    - 761 texts
    - 6.5 million words
  - Early Classics of Modern Finnish Literature
    - 13 sub-collections
    - 89 texts
    - 1.35 million words
  - Texts of Modern Finnish
    - 1 sub-collection
    - 76 texts
    - 62 105 words
  - Finnish Proverbs
    - 20 parishes
    - 72 580 proverbs

The corpora of Old Literary Finnish and Early Modern Finnish are results of many years of persistent work. The texts included in these collections are carefully selected and then digitized, either by hand or even by scanning at a later stage. The texts are marked up to the sentence level, and references to corresponding passages in the original publications are added to each sentence. Some of the texts in the corpus of Early Modern Finnish are word lists, and follow the TEI P4 dictionary structure.

The collection of the Early Classics of Modern Finnish Literature includes short stories, poems, plays and novels from the period of the beginning of modern Finnish literature. The markup goes down to the smallest structural (non-linguistic) level of the text. This may mean paragraphs in the prose texts, lines or stage descriptions in a play, or lines in a poem. For drama and poetry, new DTDs were created using the TEI Pizza Chef.

The collection of Modern Finnish is special in the sense, that the texts currently belonging to it are marked up to the word level and annotated morphologically. On the annotation scheme, see Lehtinen and Lounela (2004). Also, the text type division adds one level to the corpus structure, as the texts in the collection (at present) are divided further based on the author. At the moment the modern Finnish collection is very limited, consisting only of a collection of New Year's speeches of the presidents of the republic. In the near future certain legislative texts will be added, and we are planning to identify and encode other current texts that are not bound by the copyright laws. RILF has at its possession more texts from the twentieth century than the Kaino data service contains. These texts are arranged according to the corpus model described in this paper, and structured to the paragraph level. This material, however, is bound by copyright law, and cannot be re-published freely by RILF.

The collection of Finnish proverbs is a part of larger collection, which is digitized only partly. The collection is arranged regionally according to parishes where the proverbs were originally collected in the 1930s. The subcollections are structured according to the same DTD as the word lists and dictionaries in other collections.

7 Expandability

The corpus model of the RILF text corpus collection allows many possibilities of expanding it: by size, by function, and by information. Increasing the size of the corpus is straightforward. A compatible corpus tree (or document) can be plugged in by adding a link to it to the meta description in a suitable place in the existing tree. The new part will be readily integrated into the structure.

Expanding the corpus system by function is possible by enhancing the query interface. Exploring the XML-encoded documents with Xquery language is limited practically only by the richness of the anno-

tation expressed with the XML code. A way of exploring morphologically annotated texts is presented in Lounela (2006). This type of a summarizing analysis can be a natural part of a corpus user interface, such as the one described in this article. Linguistic summaries can be added statically, by producing them off-line and linking the reports to the existing collections in the browsing system, or dynamically, by adding queries to the searching interface.

Expanding the corpus system by information can be done by using the linking capabilities of the metadata system of the browsing interface. Above, I mentioned the possibility to link linguistic summaries or reports describing the textual properties of the collections. It is also possible to enhance the metadata system by linking the collection descriptions to external documents. The external documents may concern research that describes or uses the corpus, or is relevant to the potential corpus users in some other way.

8 Discussion

Above, I have described a corpus structure and interface that is modular, and is based on (slightly modified) de facto standard definitions belonging to the XML family. The implementation of the described interface uses open source tools. I am now going to sum up and discuss some advantages and uses of this type of a system for using corpora in linguistic work, and outline some possible future developments for such a system.

- The de facto standard -boundedness of the system makes it relatively easy to understand, implement and alter according to local needs. The existing implementation shows that a versatile user interface for the corpora can be implemented with a reasonable effort using free software.

- The modularity of the system makes it easy to add or remove corpora, sub-collections or documents, or to rearrange the corpora. Normally, parts of the system are only attached to one well-defined location in the metadata structure, which makes altering the linking straightforward.

- The integrated metadata system makes it easy to express, add, and use meta information at all levels of the corpus. The metadata can be used for browsing and searching.

- The linking system of the meta descriptions, and the use of Xquery language for the searching interface lead to expandability in many directions. New collections and documents can be plugged in, new, intelligent queries can be programmed and included, and new links, internal or external, can be added to the system in a well-defined, controlled manner.

- The searching interface can use the meta descriptions for forming new collections of already existing texts. Also, new sub-collections can be dynamically formed, regrouping the existing text documents, using existing meta information about the texts. With consistent instructions, humanistically oriented researchers can collect and annotate text materials of their own interest to be plugged in the system, and to be used in connection with previously collected materials.

- Summarizing linguistic properties with Xquery, in connection with controlled use of TEI XML structure for morphologically annotated text sets makes it possible to use text collections for text linguistic research. In Kankaanpää (2006) and Tiililä (2007) quantitative properties of text sets are examined and compared automatically in this manner as a part of text linguistic research. The techniques used in these research projects are very similar to the techniques used in the RILF corpus system.

- The corpus system could be expanded to a direction where it could be used as a tool for analyzing and recognizing text types according to frequencies of linguistic categories. On this type of research, see Biber (1988) and Saukkonen (2001).

- The mechanism of linking external documents to the metadata system of a corpus makes it possible to offer the corpus users information about research carried out using the corpus. If

this possibility was used in its full potential, the corpus could include a structured research reference database concerning the corpus and its use.

The points listed above lead to two possible (not mutually exclusive) directions of development for the corpus system. First, the system can be directed towards a research reference database where, in addition of the text materials, knowledge of relevant research is accumulated to be easily used by the researchers. Second, the system can be directed towards a general quantitative tool for text linguists. The latter would be done by expanding the searching interface to a full-blown analysis tool that offers a possibility for creating varied quantitative reports for text sets that are morphologically annotated in the standard way.

References

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Dave Beckett, Eric Miller, and Dan Brickley. 2002 *Expressing Simple Dublin Core in RDF/XML*. http://dublincore.org/documents/dcmes-rdf-xml/. The Dublin Core Metadata Initiative.

Salli Kankaanpää. 2006. *Hallinnon lehdistötiedotteiden kieli*. [Language of Administrative Press Releases.] Finnish Literature Society (SKS), Helsinki.

Stefan Kokkelink and Roland Schwänzl. 2002 *Expressing Qualified Dublin Core in RDF / XML*. http://dublincore.org/documents/dcq-rdf-xml/. The Dublin Core Metadata Initiative.

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

Mikko Lounela. 2002. Aiming Towards Best Practices in XML Techniques for Text Corpora Annotation: City of Helsinki Public Works Department - A Case Study. *Towards the Semantic Web and Web Services. Proceedings of the XML Finland 2002 Conference*. Institute for Information Technology, Helsinki.

Mikko Lounela. 2006. Exploring morphologically analyzed text material. *Inquiries into words, constraints and contexts. Festschrift in the honour of Kimmo Koskenniemi on his 60th birthday*. Gummerus, Helsinki

Wolfgang Meier et al. 2007. *Open Source Native XML Database*. http://exist.sourceforge.net/. Open Source Technology Group, Fremont, California.

Pauli Saukkonen. 2001. *Maailman hahmottaminen tekstein*. [Perceiving the world by texts.] Helsinki University Press, Helsinki.

Ulla Tiililä. 2007. *Tekstit viraston työssä*. [Texts in the work of a city department.] Finnish Literature Society (SKS), Helsinki.

DCMI visited 27.2.2007. *Dublin Core Metadata Initiative*. http://dublincore.org/. The Dublin Core Metadata Initiative.

RILF visited 27.2.2007. *Kaino - Kotuksen aineistopalvelu*. [Kaino - RILF Data Service.] http://kaino.kotus.fi/. Research Institute for the Languages of Finland, Helsinki.

TEI visited 27.2.2007 a. *TEI: Yesterday's information tomorrow*. http://www.tei-c.org/. The Text Encoding Initiative, Charlottesville, Virginia.

TEI visited 27.2.2007 b. *TEI Pizza Chef*. http://www.tei-c.org/pizza.html. The Text Encoding Initiative, Charlottesville, Virginia.

W3C visited 27.2.2007 a. *Resource Description Framework (RDF)*. http://www.w3.org/RDF/. The World Wide Web Consortium, Cambridge.

W3C visited 27.2.2007 b. *XSL Transformations (XSLT)*. http://www.w3.org/TR/sxlt. The World Wide Web Consortium, Cambridge.

W3C visited 27.2.2007 c. *Cascading Style Sheets*. http://www.w3.org/Style/CSS/. The World Wide Web Consortium, Cambridge.