

## Currently planned DECL editions:

Honkapohja, Alpo:

*A Digital Edition of MS O.1.77 Trinity College Cambridge*

A digital online edition of a late Middle English pocket-sized medical handbook, produced ca. 1460 in London or Westminster. The manuscript contains ca. 15 texts, in English and Latin. The edition will include all of MS O.1.77 and give equal attention to Middle English and Latin texts in the codex, making it the first bilingual digital edition of scientific writing in England.

Kaislaniemi, Samuli:

*The Early Letters of Richard Cocks, English Merchant (1600-1610): A digital edition*

An edition of some 100 intelligence letters written by Richard Cocks at Bayonne in France between 1603 and 1608, and sent to Thomas Wilson, secretary to Robert Cecil, in London. Cocks later worked for the *East India Company* in Japan 1613-1623, from which period letters and a journal survive and have been published.

Marttila, Ville:

*Potage Dyvers: A digital edition of a family of six late medieval culinary recipe collections*

An edition of six closely related Middle English culinary recipe collections from the 15th century. The collections are not direct copies of each other or a common ancestor, but are closely related in terms of their material. Almost all of the over 200 recipes contained in the collections appear in several members of the family, but with widely varying linguistic realisations.

The DECL project is actively looking for new contributors and collaborators  
– welcome aboard.

### Related reading:

Burnard, Lou; Katherine O'Brien O'Keefe & John Unsworth (eds.) 2006. *Electronic Textual Editing*. New York: The Modern Language Association of America.

Dollinger, Stefan. 2004. "Philological computing' vs. 'philological outsourcing' and the compilation of historical corpora: A Late Modern English test case". *Vienna English Working Papers (VIEWS)* 13 (2), 3-23.

Lass, Roger. 2004. "Ut custodiant litteras: Editions, Corpora and Witnesshood". *Methods and Data in English Historical Dialectology*, ed. Marina Dossena and Roger Lass. Bern: Peter Lang. 21-48.

Meurman-Solin, Anneli. 2001. "Structured text corpora in the study of language variation and change". *Literary and Linguistic Computing*. 16/1: 5-27.

Meurman-Solin, Anneli 2007. *Manual to the Corpus of Scottish Correspondence (CSC)*. Helsinki: University of Helsinki. Available online at <<http://www.helsinki.fi/varieng/csc/manual/>>. Accessed 6 May 2008.

Robinson, Peter 2005. "Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future?" *Digital Medievalist* 1:1 (Spring 2005). Available online at <<http://www.digitalmedievalist.org/journal/1.1/robinson/>>. Accessed 12 May 2008.

Robinson, Peter Forthcoming. "Current Directions in the Making of Digital Editions: towards interactive editions". *Ecdotica*. Prepublication version available online at <<http://www.itsee.bham.ac.uk/DistributedEditions/ecdotica.pdf>>. Accessed 12 May 2008.

*TEI Guidelines for Electronic Text Encoding and Interchange*. Available online at <<http://www.tei-c.org>>. Accessed 6 May 2008.

### Software references:

*AGTK: Annotation Graph Toolkit*. Available online at <<http://agtk.sourceforge.net/>>. Accessed 6 May 2008.

*Anastasia*. Available online at <<http://anastasia.sourceforge.net/>>. Accessed 6 May 2008.

*GATE (A General Architecture for Text Engineering)*. Available online at <<http://gate.ac.uk/>>. Accessed 6 May 2008.

*Heart of Gold*. Available online at <<http://www.delph-in.net/heartofgold/>>. Accessed 6 May 2008.

*NLTK – the Natural Language Toolkit*. Available online at <[http://nltk.org/index.php/Main\\_Page](http://nltk.org/index.php/Main_Page)>. Accessed 6 May 2008.

<teipublisher>: *A TEI Publishing system*. Available online at <<http://teipublisher.sourceforge.net/docs/index.php>>. Accessed 6 May 2008.

*Xaira (XML Aware Indexing and Retrieval Architecture)*. Available online at <<http://www.oucs.ox.ac.uk/rts/xaira/>>. Accessed 6 May 2008.

# Digital Editions for Corpus Linguistics

## Representing manuscript reality in electronic corpora

Alpo Honkapohja  
alpo.honkapohja@helsinki.fi

Samuli Kaislaniemi  
samuli.kaislaniemi@helsinki.fi

Ville Marttila  
ville.marttila@helsinki.fi

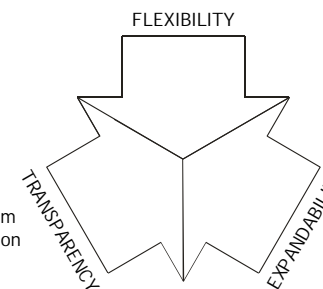
The Digital Editions for Corpus Linguistics project at the Research Unit for Variation, Contacts and Change in English  
Department of English, P.O. Box 24, FI-00014 University of Helsinki, Finland  
<http://www.helsinki.fi/varieng/domains/DECL.html>

The Digital Editions for Corpus Linguistics (DECL) project aims to create a framework for producing online editions of historical manuscripts suited for both corpus linguistic and historical research. This framework, consisting of a set of guidelines and associated tools, is designed especially for small projects or individual scholars.

A completed DECL edition will, in effect, constitute a lightly annotated corpus text. In addition to a faithful graphemic transcription of the text itself, DECL editions also contain much of the underlying manuscript reality, including features like layout and scribal annotation, together with a normalised version of the text. All of these features, encoded in standoff XML, can be used or ignored while searching or displaying the text.

The theoretical basis of the project rests on the opinion of Lass (2004) that a corpus should preserve a text as accurately as possible, convey it as flexibly as possible, and keep editorial intervention visible and reversible. Based on this, we have defined the principles of *flexibility*, *transparency* and *expandability* to serve as the guiding principles in the development of both the editions and the framework for producing them.

- use of XML markup
  - allows conversion into various existing and future formats, and the recombination of text and metadata into new documents (e.g. subcorpora)
- layered structure with customisable online interface
  - allows viewing, searching and downloading only relevant aspects of the edition
- platform-independent solutions based on the open-source principle
  - allow tools, texts and tagging to be freely downloaded and modified



- all editorial intervention indicated by markup
  - explicitly distinguished from the unemended transcription
  - transparent and reversible
- all layers of the edition accessible
  - manuscript images, raw transcript, annotation, and combinations of these
  - enables users to (re-)evaluate editorial decisions
- uniform editorial and encoding practices
  - allow new editions to be combined to form corpora
- modular architecture
  - allows new documents to be added to editions
- layered structure
  - allows new layers of annotation to be added to an edition

Figure 1. The principles of *flexibility*, *transparency* and *expandability*.

## Problems of traditional historical corpora

Problems caused by creating corpora from printed editions:

- editions often critical ones
  - do not represent authentic language use
- editorial principles vary
  - principles sometimes not made explicit
  - no comparability between editions
- orthography unreliable
  - often normalised to varying degrees
  - does help with spelling variation
- manuscript features (e.g. layout, hand changes, emendations) rarely marked
  - cannot be annotated in corpus
- copyright issues
  - publication requires clearance from publishers
- duplication of effort and errors
  - digitising is re-editing

Problems caused by conventions in the compilation and structure of corpora:

- corpora viewed as monolithic entities
  - compilation mostly limited to large or long-term projects
- closed product-like architecture
  - difficult to integrate new content
- use of corpus-specific markup
  - limited compatibility
  - limits the use of third-party tools
- little attention to presentational features
  - limited visual presentation of texts

## Solutions and improvements proposed by DECL

Solutions based on creating corpora from original manuscript sources:

- diplomatic transcript of a single witness
  - represents authentic language use
- editorial principles defined and documented by the DECL guidelines
  - constant and known practices
- original spelling preserved, including word-division and punctuation
  - words tagged with normalised forms
- manuscript features encoded into the edition using XML tagging
  - automatically included in corpus
- copyright of transcription lies with editor
  - copyright of images still an issue
- the edition serves as a corpus text
  - editing includes digitising

Solutions offered by the DECL framework:

- corpora viewed as modular and flexible
  - individuals or small projects can jointly create corpora one text at a time
- open process-like architecture
  - easy to add new annotation layers
- TEI-compatible XML markup language
  - widely supported open standard
  - easy conversion through XSLT
- advanced presentational features
  - full visual features of a digital edition

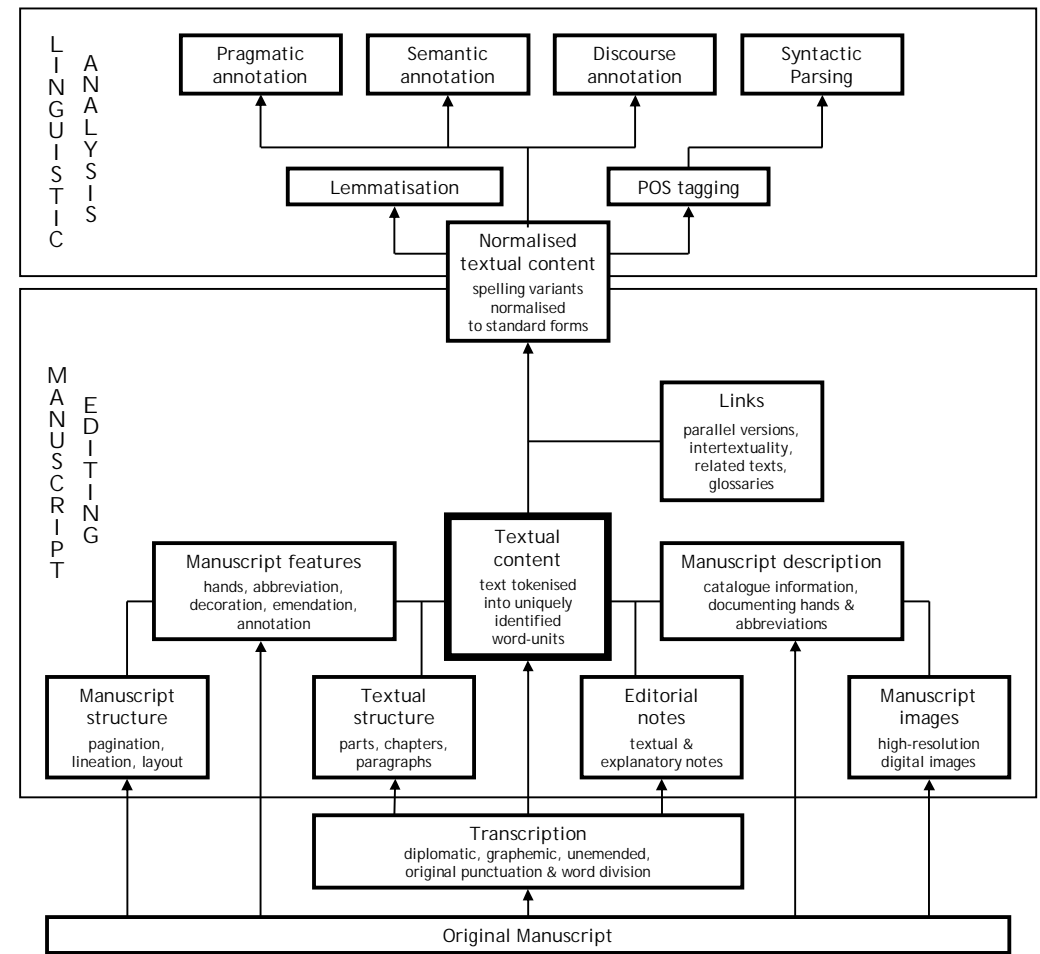


Figure 2. The highly modular structure of DECL editions allows the corpus compiler to use the normalised text as a nexus for anchoring layers of linguistic annotation to the edition without losing the connection to the underlying manuscript reality.

### TEI-compliant XML

DECL editions use XML markup based on the TEI *Guidelines for Electronic Text Encoding and Interchange*. Expressed as a modular XML schema, they define a markup language for representing the structural, visual and conceptual features of texts. DECL annotation practices will be defined by a strictly defined subset of the TEI Schema and documented in detail. The DECL guidelines will be modular in nature, and consist of a core module together with text-type-specific add-on modules.

### Compatibility

The DECL framework is designed to be as compatible as possible with existing annotation, analysis and presentation tools. To facilitate the use of 3<sup>rd</sup> party tools, a web-based interface will allow for the downloading of the edition in various formats, including the original XML representation. Furthermore, the DECL framework does not restrict the annotator to any given linguistic annotation scheme, but can be used to encode a variety of schemes, such as CLAWS, CSC, NUPOS or Penn Treebank.

### Stand-off markup

Finished DECL editions will consist of a base text document and a collection of separate *annotation layers* containing both metadata and links to external supporting data, and stored as stand-off markup anchored to the base text. This approach enables the base text to remain stable, while new layers of annotation can be added and dynamically combined with it by the user interface to create task-specific composite documents.

### Free software & Open access

The tools of the DECL project will be largely based on existing free software solutions (e.g. AGTK, Anastasia, GATE, Heart of Gold, NLTK, *teiPublisher*, Xaira). Also the DECL framework itself, together with all tools developed by the project, will be released under a suitable free license. Following the principle of transparency, DECL editions are designed to be published as free online editions that provide access to all levels of data.