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.


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 Dyers: 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 400 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.

Related reading:


Other related projects:


Digital Editions for Corpus Linguistics

A new approach to creating editions of historical manuscripts

Alpo Honkapohja
alpo.honkapohja@helsinki.fi

Samuli Kaisaniemi
samuli.kaisaniemi@helsinki.fi

Ville Marttila
ville.marttila@helsinki.fi

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.

In addition to a faithful graphemic transcription of the text itself, DECL editions will 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.

• 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 configurable interface
  - allows viewing, searching and downloading only relevant aspects of the edition
  - platform-independent solutions based on the open-source principle
  - allows tools, texts and tagging to be freely downloaded and modified

• all editorial intervention indicated by markup
  - explicitly distinguished from the unedited transcription
  - allows editors to add new layers of annotation to be added to an edition

• all layers of the edition accessible
  - manuscript images, raw transcript, annotation, and their combinations
  - enables users to (re-)evaluate editorial decisions

• documentation
  - transcription principles, editorial policies and encoding practices thoroughly documented

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.

• uniform editorial and encoding practices
  - ensure comparability of editions
  - allow editions to be combined into corpus

• 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.
Digital editions and linguistic computing

The DECL project is rooted in historical linguistics and its tradition of creating linguistic corpora - structured text databases designed for the quantitative analysis of linguistic features. The power of linguistic corpora is largely based on detailed linguistic annotation of the corpus texts and on sophisticated analysis tools. However, historical corpora in particular suffer from the consequences of "philological outsourcing" (Dollinger 2004), i.e. the compilation of corpora from printed editions, which severs the link between source document and corpus text.

Digital editions, on the other hand, have often been designed to represent manuscript reality, most commonly by including facsimile images of their source documents. Where they fall short of printed editions, which severs the link between source document and corpus text.

The DECL project aims to combine the strong points of corpora and digital editions into a new type of digital text, and create a model for editions which serve not only as editions, but also as linguistic corpora.

**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 into 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
  - 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

**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.

**Compatibility and copyright**

DECL uses TEI-compliant XML to ensure maximum compatibility with existing annotation, analysis and presentation tools.

The DECL framework will be based on existing free software solutions (such as Anastasia or teiPublisher). The framework, together with all tools developed by the project, will be published online under a suitable free license. DECL editions will also be published free online.

---

**Diagram:**

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.