

DRAFT VERSION. This paper has been submitted for publication. Please do not cite this version without permission from the DECL project (which we're likely more than happy to give – just send us an email).

\*\*\*

## **Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora**

*Alpo Honkapohja, Samuli Kaislaniemi and Ville Marttila*

Research Unit for the Variation, Contacts and Change in English (VARIENG)  
University of Helsinki

### **Abstract**

*This paper introduces a new project, Digital Editions for Corpus Linguistics (DECL), which aims to create a framework for producing online editions of historical manuscripts suited for both corpus linguistic and historical research. Up to now, few digital editions of historical texts have been designed with corpus linguistics in mind. Equally, few historical corpora have been compiled from original manuscripts. By combining the approaches of manuscript studies and corpus linguistics, DECL seeks to enable editors of historical manuscripts to create editions which also constitute corpora.*

*The DECL framework will consist of encoding guidelines compliant with the TEI XML standard, together with tools based on existing open source models and software projects. DECL editions will contain diplomatic transcriptions of the manuscripts, into which linguistic, palaeographic and codicological features will be encoded. Additional layers of contextual, codicological and linguistic annotation can freely be added to the editions using standoff XML tagging.*

*The paper first introduces the theoretical and research-ideological background of the DECL project, and then proceeds to discuss some of the limitations and problems of traditional digital editions and historical corpora. The solutions to these problems offered by DECL are then introduced, with reference to other projects offering similar solutions. Finally, the goals of the project are placed in the wider context of current trends in digital editing and corpus compilation.*

### **1. Introduction**

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 will also contain information about 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.

DECL was formed by three postgraduate students at the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki in 2007. We shared a dissatisfaction with extant tools and resources, believing that digitised versions of historical texts and manuscripts generally failed to live up to expectations. At the same time, we recognised that digitisation was time-consuming and complicated, and thus compromises had been made in the creation of digital editions and corpora. In order to alleviate these problems, we began the design of a user-friendly framework for the creation of linguistically oriented digital editions created using extant standards, tools and solutions.

The first three DECL editions will form the bases for the doctoral dissertations of the writers. Each of these editions—a Late Medieval bilingual medical handbook (Alpo Honkapohja), a family of 15th-century culinary recipe collections (Ville Marttila), and a collection of early 17th-century intelligence letters (Samuli Kaislaniemi)—will serve both as a template for the encoding guidelines for that particular text type and as a development platform for the common toolset. The editions, along with a working toolset and guidelines, are scheduled to be available within the next five years.

## 2. Theoretical and ideological orientations

Editing involves making decisions which are practical on the surface, but have underlying hermeneutic and theoretical implications (cf. e.g. Machan 1994: 2–5). When the aim is to create digital editions which encode a wide range of manuscript-related phenomena into standardised XML markup, the challenge to editorial principles is significant. The issue is further complicated by the heterogeneous target audience: historians and linguists can have widely differing assumptions about what constitutes data and how it should be presented. Consequently, it is necessary to outline the underlying theoretical orientations of the DECL project, and to place them in the context of theory and bibliographical practice within the field.

### 2.1 Artefact, Text and Context

In order to conceptualise and model the various types of information encoded into a DECL edition, we use a three-fold division of *artefact*, *text* and *context*. By *artefact* we refer to the actual physical manuscript, by *text* to the linguistic contents of the artefact, and by *context* to both the historical and linguistic circumstances relating to the text and the artefact. This division originates in the discussion of a similar categorisation in Shillingsburg (1986: 44–55), and especially in its practical application by Machan (1994: 6–7) to editing Middle English texts. The concepts of text and artefact also roughly coincide with the terms *expression* and *item* defined in Functional Requirements for Bibliographic Records (FRBR: 13).

Since DECL is concerned with what Shillingsburg calls the historical orientation of editing (1986: 19), our starting point and primary focus is the individual artefact. We see the text as a cultural product and interesting in itself, not merely as a manifestation of a work of art produced by an individual author, on which systems like FRBR tend to focus. The concept of “a work” is not a simple question, and may create more problems than it solves when dealing with texts like personal letters or a collection of anonymous culinary recipes written down in several hands. The question of authorship can also be problematic with medieval and Early Modern texts. As a result, we have decided to omit both categories, since they run the risk of making the framework too rigid to deal with non-literary historical manuscript texts.

On the other hand, our focus on texts as cultural products has led us to add the concept of *context* to represent the outside circumstances related to the production and use of the artefact and the text. Context covers the various types of cultural, social and historical background material and bibliographical information that is included in an edition. In renaming Shillingsburg’s “document” to “artefact”, we have wanted to avoid confusion and overlap with the widely accepted meaning of “document” in linguistic computing: the electronic text created by the editor.

These terms are designed to illustrate the interrelationships of the different types of features that are encoded in a DECL edition. They are meant as fuzzy rather than rigid categories, and serve to theorise how the non-linguistic aspects—including historical, codicological and bibliographical—relate to the textual whole. They serve as the foundation for a model of editing that aims to be comprehensive enough to cover all of the tasks involved in editing historical manuscripts, yet flexible enough to be adaptable to the needs of different editing projects.

### 2.2 Editorial principles

The field of historical linguistics has seen some recent discussion over what is required of an edition or corpus to be suitable for historical linguistic study (cf. i.a. Bailey 2004; Curzan and Palmer 2006; Dollinger 2004; and Grund 2006). Most vocal in his criticism of existing practices has been Lass (2004), who demands that in order to serve as valid data for the historiography of language, a digital edition or a corpus should not contain any editorial intervention that results in substituting the scribal text with a modern equivalent. He gives examples of several commonplace editorial practices, such as invisible emendations, silent expansion of abbreviations, modernisation of punctuation and word division, and attempts to construct lost archetypes based on multiple manuscript witnesses. All of these deny the reader access to information present in the manuscript original, and instead create a new artificial language variant (Lass 2004: 22). To avoid this, Lass (2004: 40) defines three criteria which he considers inviolable for a historical corpus:

- i) Maximal information preservation
- ii) No irreversible editorial interference
- iii) Maximal flexibility

While being very polemic, Lass does raise useful points and expose a number of harmful practices within historical linguistic study. It is clear that the requirements he proposes are something that compilers of editions should take into account, and therefore we have used them as a starting point, developing them further into three principles: *flexibility*, *expandability* and *transparency*. In

actual practice, these three principles influence practical considerations such as tagging, data structure, and interface design:

#### FLEXIBILITY

DECL editions seek to offer a flexible and user-friendly interface, which will allow the user to select the features of the text, artefact and context to be viewed or analysed. All editions produced within the DECL framework will build on similar logic and general principles, which will be flexible enough to accommodate the specific needs of any text type.

#### TRANSPARENCY

The user interface of DECL editions will include all the features that have become expected in digital editions. But in addition to the edited texts and facsimile images of the manuscripts, the user will also be able to access the base transcripts and all layers of annotation. This makes all editorial intervention transparent and reversible, and enables the user to evaluate any editorial decisions. In addition, the DECL framework itself will be extensively and clearly documented.

#### EXPANDABILITY

DECL editions will be built with future expansion and updating in mind. This expandability will be three-dimensional in the sense that new editions can be added and linked to existing ones, and both new documents and new annotation layers can be added to existing editions. Furthermore, DECL editions will not be hardwired to a particular software solution, and their texts can be freely downloaded and processed for analysis with external software tools. The editions will be maintained on a web server and will be compatible with all standards-compliant web browsers.

The DECL framework does, however, go further than Lass, who is primarily concerned with retaining the original scribal text. We adopt a similar position also with respect to the artefact and its context, which are treated as equally important aspects of manuscript reality and subject to the same qualitative requirements as the text itself.<sup>1</sup>

### 3. Limitations of earlier digital editions

An increasing number of digital editions of historical texts is being published online, which, for the linguist, is a mixed blessing. On the one hand access to a greater number of texts on the web obviously makes new areas of research possible, on the other, not all of the editions are amenable to linguistic enquiry. What a linguist would ideally need includes:

- i) access to the language of the original text in unadulterated form
- ii) full text searchability with sortable and refinable search results and
- iii) the possibility of defining and, preferably, extracting sub-corpora

However, digital editions range from ones providing only facsimile images to those having interfaces with most, if not all of the features listed above.

Facsimile editions are a type of digital edition common in particular to repositories—namely libraries and archives—which are primarily concerned with preservation and sustainability of digitised resources. Examples include the *Papers of Joseph Banks* at the State Library of New South Wales, but also the *Boyle Papers Online* at Birkbeck College, University of London. Editions including both facsimile images and transcripts of all or some of the texts, such as the *Auchinleck Manuscript* or the *Hooke Folio Online*, are more useful to linguists as they usually have (often limited) search functions, but these editions rarely allow users to download all the texts. Examples of editions with more elegant interfaces include the *London Provisioner's Chronicle* (diary of Henry Machyn) and the *Letters of Clemency from the Chancery of Brittany*. This last example is the best one from a linguist's viewpoint, for it not only provides the online user with facsimile images, diplomatic transcripts and indexes, but also allows the user to download the entire edition.

Commercial publications tend to have better functionality than freely available online editions. Examples of these for single works are the *Canterbury Tales* and other editions produced by Scholarly Digital Editions or Evellum, and of larger projects, the upcoming *State Papers Online*.<sup>2</sup>

The great variety among digital editions is somewhat alleviated by the fact that unlike most historical corpora (on which see below section 4.7), many of them do use TEI XML—but few of them make full use of the potential of their XML encoding. What the DECL framework aims to do is to increase comparability across the board by encouraging editors to cater to the 'linguist's needs' listed above. The framework is primarily intended for small-scale projects, which usually lack extensive funding, and which would greatly benefit from the development of more user-friendly tools and guidelines (cf. Robinson 2005).<sup>3</sup>

#### 4. Problems with traditional historical corpora

Much of the inspiration for the DECL framework comes from the shortcomings of traditional historical corpora as perceived from the point of view of a textual scholar. Most of the problems associated with using traditional historical corpora stem from the fact that because the transcription and digitisation of original manuscript texts into machine-readable form takes a lot of time and expertise, most historical corpora are based on printed editions, which “have generally not been produced with linguistic study in mind, and may not always be reliable” (Kytö et al. 2007: section 3).

##### 4.1 Use of critical editions

The most obvious problem occurs when corpora are based on critical editions which compound multiple manuscript witnesses into a single text. Compiling a corpus from these types of editions multiplies the problems inherent in them. Combining elements from several textual variants, and potentially widely differing dialectal features or scribal practices, introduces a layer of linguistic hybridity which represents the language of the editors, not of the text.

In the best case scenario, the limitations are clearly documented and acknowledged; in a more likely case, the inclusion of the text into the corpus obscures the textual nature of the edition used and thus also the unsuitability of the text for many linguistic research questions. Despite the prevalence of critical editions in the tradition of textual editing, some text types - such as letters and other unique documents - form an exception by being frequently available as single-witness editions, thus avoiding the problem of linguistic hybridity and being suitable for the study of at least morphology and syntax, as well as pragmatics (Nurmi 1999: 55).

##### 4.2 Varying editorial principles and loss of manuscript features

Another problem related to the use of editions is caused by their varying editorial principles. This problem is especially acute in corpora containing texts from both editions of varying types and from original manuscript sources. A prime example, despite all its strengths, is the important and pioneering *Helsinki Corpus of English Texts* (HC). As Kytö (1996: section 2) points out, “editorial and typographical conventions vary in different source texts (e.g. emendations can be indicated by italics, parentheses, brackets etc.)”, and “a number of ‘text level’ codes have been used to transfer the function of the convention to the computerised version, irrespective of the particular format followed in the source text”.

Although this kind of practice would on the surface seem to produce a uniform result, the format used, the amount of editorial intervention, and the degree to which various features of the original have been included can vary significantly between texts. Not many printed editions of prose texts reproduce the original layout of the text even to the level of manuscript lineation, and even fewer indicate textual details such as hand changes, scribal emendations or abbreviations. This will result in either corpus texts having a variable amount of detail encoded in them, or omitting detail from those texts that would have it. The latter phenomenon is visible in HC, which has omitted original folio and page changes, customary for most critical editions.

In either case, textual or physical features that are not recorded in all of the editions cannot be used for analysis. The worst case scenario in this respect would be a corpus that encodes the features found in each edition without any information about the principles behind the editorial decisions. Fortunately, many corpus compilers do recognise the heterogeneous nature of the corpus contents:

Editing policies vary a great deal during the 160-year history of editing medical texts, the scope being from the construction of hypothetical “originals” to faithful transcriptions. MEMT represents the “edited truth” of the underlying manuscript reality and we have reproduced the editions according to our principles [...]. Thus the texts are twice removed from their manuscript reality. (*Middle English Medical Texts* (MEMT): Introduction)

##### 4.3 Predetermined research focus

Reliance on existing editions, regardless of their editorial principles, results in another type of problem, often overlooked perhaps because of its obvious nature, namely that a corpus based on edited texts is by necessity circumscribed in its material by what has previously been considered worth editing. Textual editors tend to focus on texts considered culturally or literarily “significant”, and relying solely on editions can lead to the omission of whole categories of material. As the compilers of both the *Corpus of Early English Correspondence* (CEEC) and

*Middle English Medical Texts* (MEMT) note, this problem is not limited to the realm of literary texts but affects all genres of historical writing:

A more unexpected problem is the penchant, particularly of 19th-century editors, to edit only the letters of historically important people, and ones describing important historical events. Editors often disregarded family letters concerning everyday life, which would serve as better material for historical sociolinguistics. (Nurmi 1999: 54)

Choices made by early editors tend to define the contents of e.g. literary and linguistic histories. In language histories, the early phases of scientific writing are often ignored or passed over with few comments for the simple reason that writings in this register were not known to researchers of the time. (*Middle English Medical Texts* (MEMT): Introduction)

#### **4.4 Questionable orthography**

The use of printed editions presents several problems also on the level of the text itself. The most obvious of these, prevalent especially in older editions, is the question of orthography. Few pre-1980s editions provide detailed information about their practices concerning orthography and frequently normalise spelling—not to mention punctuation—to varying degrees. While the regularisation of spelling may help with problems related to spelling variation and automated linguistic analysis, it also means that as a rule, corpora based on printed editions cannot be used for the study of orthography or any other research questions dependant on the original spelling, as noted by the compilers of CEECS:

Particularly the older editions (ie the ones included in the CEECS) cannot be relied upon in questions of spelling, as the editors' priorities were often not linguistic but historical. Even [...] newer editions [...] [are] a less than reliable source for studies of orthography. (Nurmi 1999: 55)

#### **4.5 Copyright issues**

In addition to the aforementioned problems, relating to the integrity of the text, the use of printed editions also involves problems concerning the compilation and publication of corpora. Perhaps the most restricting of these is the problem of copyright. While historical documents (at least from the Medieval and Early Modern periods) are free of copyright, modern printed editions of them usually are not. This leaves the corpus compiler with two options: either use old, out-of-copyright editions or contact the publisher (or other copyright holder) of a more recent edition for permission to include the material in a corpus, often for a considerable fee.<sup>4</sup>

Both of these approaches have their problems. Editions from the 19th or early 20th century, which are now in the public domain, often fail to meet the standards required of reliable data for historians or historical linguists, and using them will exacerbate many of the problems mentioned above.

On the other hand, since the texts in traditional historical corpora often come from a variety of sources, obtaining permissions from all copyright holders can be a daunting task. For instance, Kytö (1996: Preface) expressly acknowledges the generosity of 38 separate persons, publishers and institutions for providing permission to include their texts in the *Helsinki Corpus of English Texts* (HC). Contacting copyright holders can be very difficult and time-consuming. The corpus compiler may encounter situations where the rights have moved from one holder to another or where the institution holding them has ceased to be operational, and in the end, the current holder may or may not grant them (see e.g. *Middle English Medical Texts* (MEMT): introduction; Nurmi 1999: 56).

#### **4.6 Duplication of effort**

Two more problems that stem from using printed editions in compiling corpora are the duplication of effort and an increased probability of errors. Producing an edition of manuscript material in whatever form involves a significant amount of work. If the edition is published in printed form and used as a source for a corpus, the compiler will need either to key in the whole text or use OCR in order to digitise it. Both of these methods require at least some degree of proofreading and are likely to introduce new errors into the text. This kind of perceived waste of effort was actually one of the key issues in forming the DECL project: how could we ensure that editions would be immediately useable as corpus texts without a significant amount of additional work.

#### 4.7 Problematic corpus conventions

Traditionally, corpora have been viewed as monolithic entities—collections of texts that are compiled, digitised and annotated, and when all the stages are finished, released as a whole.<sup>5</sup> As a result, large or otherwise work-intensive corpora can spend years as ‘work-in-progress’, being generally unavailable to the scholarly community even if significant parts of them are already finished. Furthermore, this view of corpus compilation as a large undertaking involving a huge mass of texts can easily discourage small projects and individual scholars from compiling corpora, because it would take too long to compile a corpus of sufficient size.

Once a corpus is finished, it is not commonplace to make provisions for including new material. There have been several updated or expanded versions of earlier corpora,<sup>6</sup> but even they have mostly taken the form of new individual and closed products. DECL aims to provide means to add new content, either in the form of new texts (‘horizontal expansion’), additional annotation (‘vertical expansion’) or supporting background material.

The requirements posed by this kind of expandability also reveal another problematic property of many corpora, namely the use of corpus- or project-specific tagging and encoding practices, often developed for the needs of one specific corpus. Although there are some accepted and established principles and ways of encoding corpus material, the situation in the case of corpora is far from the optimistic view that seems to prevail in the field of digital humanities:

Gone, too, are the days when every individual or project invented codes, systems, or symbols of their own to identify special features, and any character that could not be represented in ASCII had to be recoded in some arcane form. (Deegan and Tanner 2004: 493–494)

This seems to be mainly a historical development. Many corpora have borrowed their encoding and markup practices from earlier corpora and adapted them to their own use.<sup>7</sup> This kind of variance limits the development and use of common tools and the convertibility of corpora from one format to another. The situation is somewhat surprising, considering that standards for the electronic encoding of textual data, most notably the Text Encoding Initiative (TEI) have been around for almost two decades (the first version of the TEI Guidelines was published in 1990). There are some historical corpora that use a version of the TEI Guidelines, such as *The Lampeter Corpus of Early Modern English Tracts*, but use of the Guidelines seems to be significantly more common in other branches of digital humanities than in corpus linguistics.

#### 4.8 Shallow representation of manuscript reality

Historical corpora are often characterised on a two-dimensional scale as long or short and thin or fat, the first reflecting their diachronic scope and the second the extent of their synchronic coverage (cf. Rissanen 2000). Comparatively less attention has been paid to a third dimension, depth, which could be defined as the extent to which the corpus represents the various features of the original texts. This dimension is especially relevant in the case of materials with limited availability, such as historical manuscripts. A deeper representation helps to widen the applicability of the corpus to different types of research, which is important for specialised corpora that run the risk of becoming marginal if their applicability is further limited by design or compilation choices.

Moreover, in contrast to digital editions, most linguistic corpora have given little attention to the visual presentation of text, being oriented towards linguistic analysis. This, together with the limited search and analysis tools provided by most digital editions, has created a wide but unnecessary rift between these two types of digital resources, which at their heart have much in common and could both benefit immensely from closer integration with each other.

### 5. Key features of the DECL framework

As a response to these problems and driven by the theoretical and ideological orientations described above, the DECL framework has been designed to overcome the limitations and combine the benefits of both digital editions and traditional historical corpora. Most of the individual features described below are not unique to DECL but are evidenced by various other corpus and digital editing projects. The aim of the DECL framework is to learn from the example of these projects and to bring together their best aspects while simultaneously avoiding as many of the abovementioned problems as possible.

## **5.1 Faithful representation of original texts**

Since the DECL framework is intended for producing digital editions of historical texts, one of its primary objectives must be the definition of clear and consistent editorial principles. Being oriented primarily (though not exclusively) towards producing editions useful for corpus linguistics, the emphasis must be on representing authentic language use. The need for more linguistically-oriented editions that “aim at reproducing the original manuscripts more faithfully than critical or eclectic editions do” (Kytö et al. 2007: section 3) has been widely acknowledged in recent years. This has also affected the compilation principles of many recent corpus projects, such as the *English Witness Depositions 1560–1760: An Electronic Text Edition* (EWD) project at the University of Uppsala, the *Middle English Grammar Corpus* (MEG-C) at the University of Stavanger, *A Linguistic Atlas of Early Middle English* (LAEME) and *A Linguistic Atlas of Older Scots* (LAOS) at the University of Edinburgh, *The Corpus of Early Ontario English* (CONTE) at the University of British Columbia, *A Corpus of Middle English Scientific Prose* (ACOMESP), a collaboration between the University of Málaga and the University of Glasgow, and the *Corpus of Scottish Correspondence* (CSC) at the University of Helsinki.<sup>8</sup>

The editors of EWD introduce the concept of a “linguistic edition” and define it as an edition where “the language of the original manuscript text is not normalised, modernised, or otherwise emended”, but “the manuscript is reproduced as closely as possible in transcription” (Kytö et al. 2007: section 3). Similarly, the compilers of the MEG-C aim “to record what is visible in the manuscript, rather than giving editorial interpretations” (Stenroos and Mäkinen 2008: 14), reproducing the text “at what might be called a rich diplomatic level” (Stenroos and Mäkinen 2008: 7). This type of linguistic edition is essentially what lies also at the heart of a DECL edition: a diplomatic transcription of an individual manuscript witness, representing a sample of authentic language use.<sup>9</sup> In the case of DECL and both of the abovementioned projects, this also entails the use of original manuscripts as the source of the edition or corpus, although microfilms and digital reproductions can be used as an aid in the editing process.

Editions produced using the DECL framework will preserve the orthography of the original manuscript down to graphemic level without normalising either spelling or punctuation. The guidelines also aim at the preservation of the original word-division, but since the word-spacing of manuscript texts is not always reproducible in digital format, editorial judgement of whether two words are separated by a space will be required in unclear cases. While preserving the original orthography, the DECL framework will also provide tools and guidelines for annotating every word token of the original text with its normalised form, facilitating searches and automated analysis of the text with tools developed for PDE.

Since the DECL framework places equal emphasis to the levels of text, artefact and context, the scope of faithful representation extends beyond the strictly textual level. DECL editions will try to represent the physical layout and appearance of the text on the manuscript page—ideally both as machine-readable tagging and in facsimile images—and provide a description of the cultural and historical context of the text.

Another aspect of faithful representation, which has traditionally been associated with digital editions rather than corpora, is the visual representation of textual and palaeographical features of the text. DECL editions will have an online interface which will be customisable in two senses. Firstly, the editors of individual DECL editions will be able to choose which features will be implemented in their edition and, since all tools developed for the DECL framework will be open source, even program new features. Secondly, the interface will enable the user to choose the features of the text to be viewed, downloaded or included in the analysis. In addition to visual presentation and browsing, the interface will also offer corpus search and analysis functions and the ability to download the texts in various formats.

## **5.2 Edition = corpus**

As pointed out above, one of the central ideas behind the DECL project is to combine the strengths of digital editions and linguistic corpora into a single multipurpose resource. Considering that many digital editions and all historical corpora are essentially digital transcripts of text whose production involves many overlapping tasks, there have been surprisingly few attempts to combine them. While it is true that many digital editions have rudimentary search tools and some corpora provide ways of visually representing the corpus texts, only a few projects attempt to create editions that would serve as corpora straight out of the box. There are some important predecessors; two examples of projects with similar aims are the EWD and ACOMESP already mentioned above. The editors of the EWD emphasise that their edition “will be geared to facilitate advanced computer searches” and that they “combine our philological and editorial aims with

principles of modern corpus compilation, striving at a new type of text edition that will also serve as a computerised corpus” (Kytö et al. 2007: section 5). ACOMESP in turn offers a web interface that allows viewing facsimiles and transcriptions side by side, and conducting corpus searches on the texts. The project also benefits from being able to use high quality facsimiles from the Hunter collection of the library of the University of Glasgow.

What, then, are the basic requirements—in addition to the faithful representation discussed above—of an edition that can be used as (part of) a corpus? The most obvious requirement is for it to include machine-readable, i.e. digital transcripts of the source texts. Next, it must be possible to perform text searches on them, preferably with support for regular expressions (or at least wildcards). This second requirement can be fulfilled either by including a suitable search engine in the interface or by allowing the text of the edition to be extracted in a format that is usable by external corpus tools—or, ideally, by both methods.

The elimination of the rift between an edition and a corpus also means that all of the textual and codicological features encoded in a DECL edition are automatically available in the corpus without need for further encoding, and all linguistic metadata added to a corpus are also available for users of the edition.

### 5.3 Modular and layered architecture

Being aimed especially at a community of small projects and individual scholars, the DECL framework promotes a view of corpora not as monolithic and closed text collections but as modular and flexible networks of texts, whose production can thus be distributed both in time and place.<sup>10</sup> In practice this means that by following the guidelines and practices defined by the DECL framework, independent scholars or projects can produce and release ‘mini-corpora’ or even individual texts, which can then be joined together into larger corpora and further supplemented with new texts. A similar process-like approach to corpus compilation has been adopted by the MEG project, although within a more traditional version paradigm.<sup>11</sup> Releasing the corpus before it is “finished” not only allows the scholarly community to benefit immediately from what has been accomplished so far, but also avoids limiting the potential size of the corpus: theoretically, new texts could be added until all known texts have been included.

The DECL guidelines have also been designed to allow for the addition of new layers of annotation to existing texts. This is made possible by the use of *standoff annotation*, where the annotation layers are maintained separate from the base text and linked to it by means of uniquely identified word tokens. These annotation layers are not limited to traditional linguistic annotation, but can contain any kind of ancillary information relating to the text.

This means that all editorial intervention and interpretation is not only indicated by markup, but also physically separated from the base text, rendering it transparent and easily reversible. While the use of annotation layers allows the user to focus on only the selected aspects of the text, they are also persistently linked together and can be freely accessed at any time. By allowing for the addition of new annotation layers to the text without changing the base text, the layered architecture not only ensures the stability of the base text, but also allows for the creation of mutually exclusive annotation layers.

In terms of corpus compilation, this means that corpora can be created simply by defining an annotation layer linking a set of texts together. The corpus compiler can also provide individual texts with descriptive attributes which allow the user of the corpus to dynamically define subcorpora. Furthermore, the texts included in the corpus can be analysed using external annotation tools, temporarily ignoring any annotation layers not relevant to the analysis. The results of this analysis can then be detached from the text and converted into a new annotation layer to be shared with others.

To facilitate the automatic linguistic analysis of DECL editions, the framework calls for the inclusion of an annotation layer containing normalised forms for every word token, eliminating—or at least alleviating—the problem of spelling variation inherent in historical corpora.<sup>12</sup>

Figure 1 below illustrates the structure of a richly annotated DECL edition that has been included in a corpus and analysed for various linguistic features, along with the division of labour between the manuscript editor and the corpus linguist.

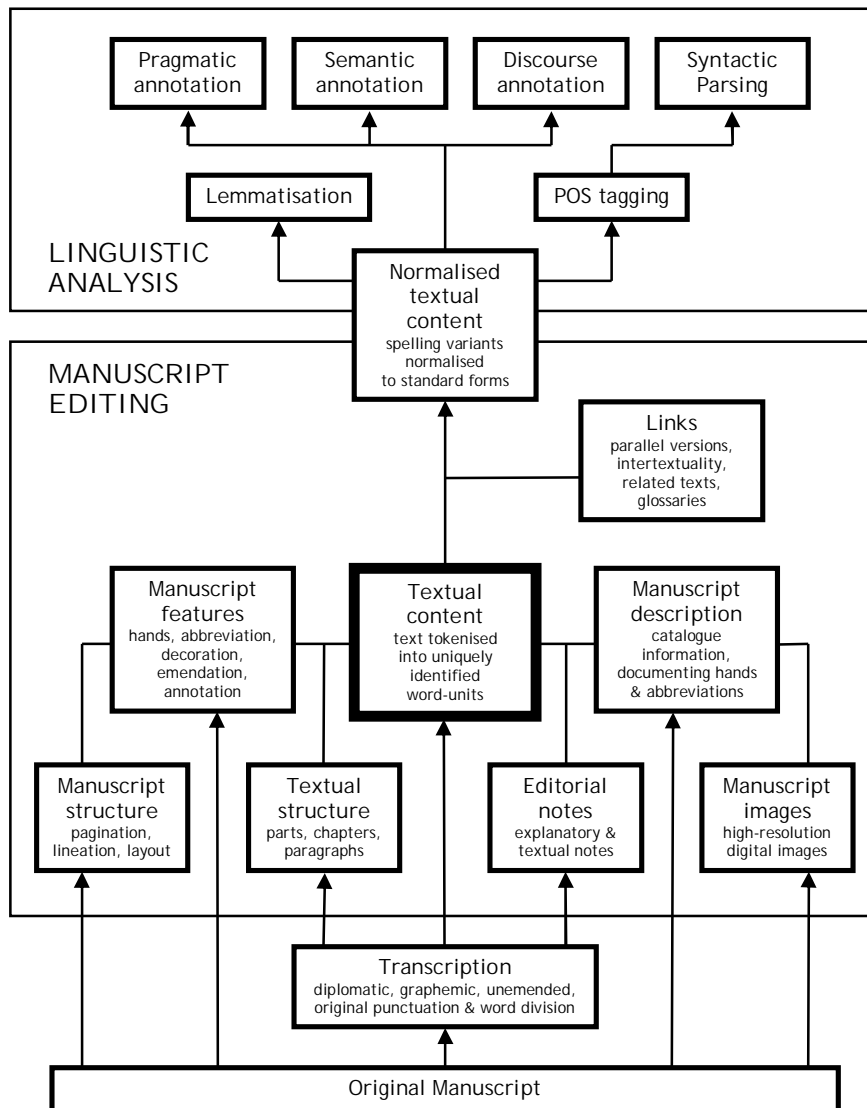


Figure 1. The conceptual structure of a DECL edition.

#### 5.4 The virtues of standards

As pointed out earlier, the field of digital humanities has seen much work in the creation of encoding and markup standards in the last decade or so. Perhaps the most significant effort in providing standard forms of textual markup has been the Text Encoding Initiative (TEI). Currently in their fifth public version (P5), the XML-based TEI Guidelines have been adopted by a large number of projects within the field of digital humanities, including the British National Corpus (BNC) and even some historical corpus projects, such as *The Corpus of Northern English texts from Old to Early Modern English* at the University of Seville.<sup>13</sup> Expressed as a modular XML schema, the Guidelines define a markup language for representing the structural, visual and conceptual features of texts.

The DECL framework is based on the TEI Guidelines, and the DECL editorial guidelines will be a strictly defined subset of the TEI schema, documented in detail. This means that any edition produced according to the DECL guidelines is automatically TEI-conformant and thus compatible with any TEI-compatible tools. Since the TEI and thus the DECL guidelines are valid XML definitions, more generic XML tools can also be readily used with DECL editions. Conversely, any tools produced within the DECL framework will also be usable (or can be modified to be so) within other TEI-compatible projects.

From a more technical viewpoint, XML brings several benefits. First of all, XML readily supports the kind of modular approach described above, and makes a clear distinction between the textual content and the markup, consisting of a defined set of elements, which can be described by assigning values to their attributes. Furthermore, XML markup provides the added advantage of using XSLT (eXtensible Stylesheet Language: Transformations) to manipulate and transform the content of documents, either to create new XML documents from the contents of the edition or to

convert it into other markup formats. This will enable DECL editions to be used with various existing annotation, analysis and presentation tools. Furthermore, the XML markup used by 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.

This makes the approach of DECL subtly but fundamentally different from that taken by other related projects, such as MEG-C (Stenroos and Mäkinen 2008: 15) or EWD (Kytö et al. 2007: section 2). Instead of releasing separate versions for different purposes (e.g. reading and linguistic searches), custom representations are created dynamically from the base XML based on the user's selections. This helps to maintain the link between all representations and the original data, meaning that for example any search results found using normalised forms of the words remain linked not only to the original forms but to all of the formatting and background information pertaining to them.

While the XML definition and the TEI guidelines have largely standardised the technical aspects of encoding text, the aim of the DECL Guidelines is to go further and to use these standards as a basis for defining and documenting a set of editorial principles and practices. This will eliminate the problem of varying editorial principles, discussed above, and allow DECL conformant editions to be used together and combined into corpora.

### 5.5 Fundamental freedom

Since the DECL project is committed to the principles of open access and open source software, all of the tools and documentation of the DECL framework will be released following these principles as far as possible.<sup>14</sup> The project intends to both make use of existing open source software projects and adapt them to its needs, and develop new custom solutions for those needs that have not yet been met by existing solutions. All tools will be developed to be platform-independent and as flexible as possible.

Naturally, these principles will also be extended to any editions produced using the framework. Using original manuscripts as sources provides DECL editions freedom from external copyright: the copyright of the transcript resides with the transcriber. In order to avoid copyright issues between transcribers, editors and corpus compilers, and to allow DECL editions to be freely used in corpora, the framework requires that DECL editions be published under a suitable open access license. A similar approach has been taken by the MEG-C corpus, which is distributed under the Creative Commons Attribution-Noncommercial-Share Alike (*by-nc-sa*) license, giving the users freedom to not only use the corpus as it is, but also to create and publish derivative works under the same license, provided that the original work is credited to its authors and the derivative work is not distributed commercially.<sup>15</sup> This particular license is also the strongest candidate considered for publishing DECL editions.

This freedom extends also to the internal workings of the edition: in keeping with the idea of transparency, all layers of the edition from the base transcript to the various levels of annotation will be accessible for viewing, searching and downloading. This will not only ensure the reusability of previously created resources, but also enable the user to evaluate any editorial decisions.

Although using open access transcriptions of original sources solves the problem of copyright for the texts, the copyright of manuscript images remains a problem. Since most manuscript repositories<sup>16</sup> reserve the right to produce digital reproductions of their collections and charge significant fees for these reproductions, small projects in particular may be hard-pressed to obtain digital facsimiles even for their own use. Furthermore, since the repository that produced the reproductions owns the copyright for them, they cannot be freely published under an open access license. The only way to get around this problem is to work with repositories and persuade them to either digitise the manuscript material and to publish them under an open access license, or to allow scholars to photograph manuscript material themselves.

With regard to corpus compilers, the DECL framework seeks to liberate corpus compilers from the chains of 'what has been edited' and enable them to add texts from original sources with reasonable effort, effectively becoming digital editors themselves. It is clear that the viability of this depends on both the nature of the material and the text-scholarly competence of the scholars concerned. Yet while the DECL framework can offer only limited assistance in the textual scholarship required for editing original manuscript texts, it will provide a thoroughly documented markup for recording the features of the manuscript text, detailed guidelines on the various steps involved in creating a digital edition, and tools to facilitate and even automate many of the steps involved in turning a base transcript into a finished digital edition.

## 6. Conclusion: Working towards mutual goals

We wrote above that DECL was triggered by a dissatisfaction with existing digital resources, and have argued that a more systematic effort should be made in the creation of digital resources of historical documents in order to increase their accessibility, usability and versatility. Similar concerns have been voiced by others, linguists and historians alike, as well as by archivists and other scholars. In the manual of the *Corpus of Scottish Correspondence* (CSC), Anneli Meurman-Solin writes that:

[T]he fourth generation of corpora will combine three important properties. Firstly, we define language-external variables rigorously, benefiting from information provided by various interdisciplinary forums. Secondly, we see corpora as consisting of sub-corpora that are defined ... in reference to degrees of validity and relevance as regards their usefulness for the study of a specific research question. Thirdly, instead of marketing corpora as completed products, we see the compilation as an ongoing process, and therefore view expansion and revision as inherent characteristics of this work. (Meurman-Solin 2007, section 2.1.1)<sup>17</sup>

Meurman-Solin's second point is one pertinent to this age of web-based corpora used for studying PDE. Yet such an approach is becoming feasible for historical linguistics as well, as shown by Hendrik De Smet's *Corpus of Late Modern English Texts* (CLMET), which he has compiled from sources already available online:

[T]he corpus can be extended or reduced at wish, and similar—though not necessarily identical—corpora can be compiled without much effort by anyone [...] The corpus presented here is what I consider an acceptable and useful offshoot of a continual attempt to open up the rich resources of the Internet to historical linguistic research. (De Smet 2005: 70)

Still, the CLMET is closer to a traditional historical corpus than the CSC, in that its sources are digitised versions of editions of Late Modern English texts, while the CSC is based on manuscripts. But as mentioned above, one of the aims of the DECL project is to eventually enable the creation of historical corpora in a similar fashion to the CLMET, based on a large number of DECL-compliant digital editions of historical documents. This objective is not a new idea, and has been dubbed a "textbase approach" to using digitised resources (Vanhoutte and Van den Branden forthcoming).

In short, the aim is to make online resources into multi-functional databases by encouraging their creation according to defined standards. As Vanhoutte and Van den Branden (forthcoming, section 10) put it, "from a rich textbase of encoded ... material ... various derived products [can] be extracted and realised, such as scholarly editions, reading texts, indexes, catalogues, calendars, regests, polyfunctional research corpora etc". The textbase approach works in tangent with the concept of "distributed" production: spreading the workload of a project by opening it to other scholars (as described above in section 5.3). Such collaboration would ultimately lead to shared online resources not entirely unlike Wikipedia (and other Wikimedia resources), but created and moderated by scholars for (primarily) scholarly purposes. These aims require collaboration at a high level, but fortunately such initiatives exist: one, for markup, is the aforementioned Text Encoding Initiative (TEI); another, for general architecture, is the Distributed Editions initiative led by the Institute for Textual Scholarship and Electronic Editing at the University of Birmingham.

The aims of DECL are much the same as those of the Distributed Editions initiative: to create versatile digital resources by adhering to agreed standards, by allowing other scholars access to improve these resources, and by helping to create multidisciplinary shared online resources. In other words, we, too, are working towards "a federated model of scholarly tools and materials on the internet", as it is phrased on the Distributed Editions website (<http://www.itsee.bham.ac.uk/DistributedEditions/summary.htm>). While these theoretical goals may sound highly optimistic, on a practical level DECL hopes to participate primarily by creating more editions of previously unedited historical manuscripts, ensuring that all of them are suited for linguistic study.

For more and up-to-date information, please visit the DECL website at <http://www.helsinki.fi/varieng/domains/DECL.html>.

## Notes

*Work done for the DECL project has been funded by the Research Unit for Variation, Contacts and Change in English (VARIENG), a Centre of Excellence funded by the Academy of Finland, and the Finnish Cultural Foundation.*

- 1 Lass does acknowledge the importance of the features of the *artefact* to some degree by mentioning the potential significance of retaining punctuation and manuscript page layout (2004: 36).
- 2 The *State Papers Online* is arguably not an edition. However, while the features of its interface are similar to those of resources like *Early English Books Online* (EEBO) and *Eighteenth-Century Collections Online* (ECCO), its scope is more strictly defined. The *State Papers Online* is not to be confused with the admirable but lo-fi effort of the *State Papers Project*, a freely available edition of a part of the same material.
- 3 The *Letters of Clemency from the Chancery of Brittany* is an exception, being a simple but highly usable and versatile digital edition along the lines encouraged by the DECL project. Yet it shows particularly well what can be done with reasonable effort, and what functionalities all digital editions could have. The *Digital Archive of Letters in Flanders* (DALF) project editions produced by the Centre for Scholarly Editing and Documents Studies (CTB) at Ghent also contain all these functionalities, yet are unfortunately not available online (Vanhouette, p.c.).
- 4 For examples of problems related to copyright in the context of corpus compilation, see e.g. *Middle English Medical Texts: Introduction and Nurmi 1999: 56*.
- 5 The *Middle English Grammar Corpus* (MEG-C) is a welcome exception, as it is intended to be published in incremental parts as the work progresses. See section 5.3 below, and note 11.
- 6 For example the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME) is built on the Middle English part of the HC, and has itself seen a second edition (PPCME2). An exception to this tendency is the *Corpus of Scottish Correspondence* (CSC), see below section 6.
- 7 Examples of this include the *Middle English Medical Texts* (MEMT) corpus and the *Corpus of Early English Correspondence* (CEEC), which both use a markup system based on that of the *Helsinki Corpus of English Texts* (HC).
- 8 It is interesting—and to some degree indicative of the intimate relationship between historical corpus compilation and digital editing—that while the MEG-C and the *English Witness Depositions 1560–1760* are quite similar in their aims and methods, the former is described as a corpus and the latter as an electronic text edition.
- 9 This emphasis on individual manuscript witnesses does not preclude multi-text editions. The DECL framework will include provisions for producing editions of several manuscript versions of a text (or several closely related texts) and presenting them as a parallel text edition, enabling the comparison and analysis of the variation between versions.
- 10 Some of the difficulties involved in distributing the tasks of editing and corpus compilation between two completely separate projects without a common framework are exemplified by the interrelationship of *The Proceedings of the Old Bailey* and the *Old Bailey Corpus* as described by Huber (2007).
- 11 The first version of the MEG-C corpus, containing roughly a third of the base texts, has already been published. New versions will be released as more texts get added, approximately every six months (Stenroos and Mäkinen 2008: 2).
- 12 The inclusion of the normalisation of the text in the editing phase instead of the corpus compilation phase is based on the assumption that the editor of a historical text is usually more familiar with both the individual text and its linguistic conventions than the corpus linguist, who is dealing with a larger selection of potentially very different texts.
- 13 The MEG project also initially considered adopting TEI XML P5 as the annotation format, but due to reasons of convenience and compatibility opted at least initially for an encoding system based on that developed for LAEME (Stenroos and Mäkinen 2008: 6). According to Mäkinen (pers. comm.), moving over to XML at some later stage has not been ruled out and the encoding system used by the project has been kept such that it can be easily converted to XML at a later date.
- 14 The TEI Guidelines themselves are available under the terms and conditions of the GNU General Public License (version 2, <<http://www.gnu.org/licenses/old-licenses/gpl-2.0.html>>), which means that the DECL guidelines will also need to be published under a compatible free software license.
- 15 For an explanation of the terms of this license, see <<http://creativecommons.org/licenses/by-nc-sa/3.0/>>.
- 16 With the notable exception of the National Archives of the United Kingdom, which allows researchers to take digital photographs of their archival material.
- 17 Meurman-Solin's properties for fourth-generation corpora are much the same as the DECL principles of transparency, flexibility and expandability, combined with the cultivation of international standards and retaining the link to the manuscript reality. See also sections 2 and 5 above.

**Editions, corpora and other related projects**

- The Auchinleck Manuscript*. <<http://www.nls.uk/auchinleck>>. Accessed 12 August 2008.
- The Boyle Papers Online*. <[http://www.bbk.ac.uk/boyle/boyle\\_papers/boylepapers\\_index.htm](http://www.bbk.ac.uk/boyle/boyle_papers/boylepapers_index.htm)>. Accessed 12 August 2008.
- British National Corpus (BNC)*. <<http://www.natcorp.ox.ac.uk>>. Accessed 18 August 2008.
- Caxton's Canterbury Tales: The British Library Copies*. 2003. Barbara Bordalejo (ed.). CD-ROM. Birmingham: Scholarly Digital Editions.
- The Centre for Scholarly Editing and Document Studies (Centrum voor Teksteditie en Bronnenstudie – CTB) at the Royal Academy of Dutch Language and Literature (KANTL) in Ghent, Belgium. <<http://www.kantl.be/ctb>>.
- Corpus of Early English Correspondence (CEEC)*. 1998. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of English, University of Helsinki. Description available at <<http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html>>.
- Corpus of Early English Correspondence Sampler (CEECS)*. 1998. Compiled by Jukka Keränen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg at the Department of English, University of Helsinki.
- The Corpus of Early Ontario English, 1776–1899 (CONTE)*. Being compiled by Stefan Dollinger at the University of British Columbia. See Stefan Dollinger (2006), 'Oh Canada! Towards the Corpus of Early Ontario English', in: Antoinette Renouf and Andrew Kehoe (eds.) *The Changing Face of Corpus Linguistics*. Amsterdam: Rodopi.
- Corpus of Late Modern English Texts (CLMET)*. <<http://perswww.kuleuven.be/~u0044428/clmet.htm>>. Accessed 18 August 2008.
- A Corpus of Middle English Scientific Prose (ACOMESP)*. Being compiled at the University of Málaga in collaboration with the University of Glasgow by Antonio Miranda García, David Moreno Olalla, Graham D. Caie, Javier Calle Martin, María Laura Esteban Sequra, Nadia Obegi Gallardo, Santiago González Fernández Corugedo and Teresa Marqués Aguado. <<http://hunter.filosofia.uma.es/manuscripts>>. Accessed 15 August 2008.
- The Corpus of Northern English texts from Old to Early Modern English*. Being compiled at the University of Sevilla. Described in Gabriel Amores Carredano, Julia Fernández Cuesta and Luisa García-García (2008), 'Elaboration of an electronic corpus of northern English texts from Old to Early Modern English', paper presented at the Sixth International Conference on Middle English, 24–26 July 2008.
- Corpus of Scottish Correspondence, 1500–1730 (CSC)*. Being compiled at the University of Helsinki by Anneli Meurman-Solin.
- Creative Commons. <<http://creativecommons.org/>>.
- DALF: *Digital Archive of Letters in Flanders*. <<http://www.kantl.be/ctb/project/dalf>>. Accessed 12 August 2008.
- Digital Editions for Corpus Linguistics (DECL)*. <<http://www.helsinki.fi/varieng/domains/DECL.html>>.
- The Distributed Editions initiative. Description available at <<http://www.itsee.bham.ac.uk/DistributedEditions>>. Accessed 12 August 2008.
- Early English Books Online (EEBO)*. Available to subscribers at <<http://eebo.chadwyck.com/home>>. Accessed 13 August 2008.
- Eighteenth Century Collections Online (ECCO)*. Available to subscribers at <<http://galenet.galegroup.com/servlet/ECCO>>. Accessed 13 August 2008.
- English Witness Depositions 1560–1760: An Electronic Text Edition (EWD)*. Being compiled by Merja Kytö, Peter Grund and Terry Walker. Description available at <<http://www.engelska.uu.se/witness.pdf>>. Accessed 18 August 2008.
- Evellum. <<http://www.evellum.com>>. Accessed 12 August 2008.
- Helsinki Corpus of English Texts (HC)*. 1991. Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Description available at <<http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html>>.
- Hooke Folio Online*. <<http://webapps.qmul.ac.uk/cell/Hooke/Hooke.html>>. Accessed 12 August 2008.
- The Institute for Textual Scholarship and Electronic Editing. <<http://www.itsee.bham.ac.uk>>.
- The Lampeter Corpus of Early Modern English Tracts*. Manual available at <<http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM>>.
- Letters of Clemency from the Chancery of Brittany*. <<http://nicole.dufournaud.net/remission>>. Accessed 12 August 2008.
- Letters of William Herle Project*. <<http://www.livesandletters.ac.uk/herle/index.html>>. Accessed 12 August 2008.
- A Linguistic Atlas of Early Middle English, 1150–1325 (LAEME)*. 2007. The University of Edinburgh. Compiled by Margaret Laing and Roger Lass. <<http://www.lel.ed.ac.uk/ihd/laeme/laeme.html>>. Accessed 8 August 2008.
- A Linguistic Atlas of Older Scots, 1150–1325 (LAOS)*. The University of Edinburgh. Compiled by Keith Williamson. <[http://www.lel.ed.ac.uk/ihd/laos1/laos1\\_frames.html](http://www.lel.ed.ac.uk/ihd/laos1/laos1_frames.html)>. Accessed 18 August 2008.

- A London Provisioner's Chronicle, 1550–1563*, by Henry Machyn: *Manuscript, transcription, and modernization*. Edited by Richard W. Bailey, Marilyn Miller and Colette Moore. <<http://quod.lib.umich.edu/m/machyn>>. Accessed 12 August 2008.
- Middle English Medical Texts* (MEMT). 2005. Taavitsainen Irma, Päivi Pahta and Martti Mäkinen (eds.). CD-ROM. Amsterdam: John Benjamins.
- Middle English Grammar Corpus* (MEG-C). 2008. Version 1.0. University of Stavanger. Compiled by Merja Stenroos, Martti Mäkinen, Simon Horobin, Jeremy Smith. <[http://www.uis.no/research/culture/the\\_middle\\_english\\_grammar\\_project](http://www.uis.no/research/culture/the_middle_english_grammar_project)>. Accessed 6 August 2008.
- Old Bailey Corpus* (OBC). Coordinated by Magnus Huber. Description available at <<http://www.uni-giessen.de/oldbaileycorpus/index.php>>. Accessed 14 August 2008.
- Papers of Sir Joseph Banks*. <<http://www2.sl.nsw.gov.au/banks/>>.
- Penn-Helsinki Parsed Corpus of Middle English* (PPCME2). 2000. 2nd edition. Anthony Kroch and Ann Taylor (eds.). Description available at <<http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/>>. Accessed 14 August 2008.
- The Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913*. Old Bailey Online. <<http://www.oldbaileyonline.org>>. Accessed 21 August 2008.
- Scholarly Digital Editions. <<http://www.sd-editions.com>>. Accessed 12 August 2008.
- State Papers Online*. Forthcoming. <<http://gale.cengage.co.uk/statepapers>>. Accessed 12 August 2008.
- State Papers Project*. Coordinated by Helen Good. <<http://www.sp12.hull.ac.uk>>. Accessed 12 August 2008.
- Text Encoding Initiative* (TEI). <<http://www.tei-c.org>>.

## References

- Bailey, Richard W. (2004), 'The need for good texts: The case of Henry Machyn's Day Book, 1550–1563', in: Anne Curzan and Kimberly Emmons (eds.), *Studies in the history of the English language II: Unfolding conversations (Topics in English Linguistics 45)*. Berlin and New York: Mouton de Gruyter. 217–228.
- Curzan, Anne and Chris C. Palmer (2006), 'The importance of historical corpora, reliability, and reading', in: Roberta Facchinetti and Matti Rissanen (eds.) *Corpus-based Studies of Diachronic English*. Bern: Peter Lang. 17–34.
- Deegan, Marilyn and Simon Tanner (2004), 'Conversion of Primary Sources', in: Susan Schreibman, Ray Siemens and John Unsworth (eds.) *A Companion to Digital Humanities*. Malden: Blackwell Publishing.
- De Smet, Hendrik (2005), 'A corpus of Late Modern English texts'. In *ICAME Journal*, 29: 69–82. Available at <<http://icame.uib.no/ij29>>.
- 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.
- Functional Requirements for Bibliographic Records* (FRBR). Final Report. IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Available at <<http://www.ifla.org/VII/s13/frbr/frbr.htm>>. Accessed 15 August 2008.
- Grund, Peter (2006), 'Manuscripts as sources for linguistic research: A methodological case study based on the Mirror of Lights', *Journal of English Linguistics*, 34: 105–125.
- Huber, Magnus (2007), 'The Old Bailey Proceedings, 1674–1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English', in: Anneli Meurman-Solin and Arja Nurmi (eds.) *Annotating Variation and Change (Studies in Variation, Contacts and Change in English 1)*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. <<http://www.helsinki.fi/varieng/journal/volumes/01/huber>>. Accessed 10 July 2008.
- Kytö, Merja (comp.) (1996), *Manual to the Diachronic Part of The Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. 3rd ed. Helsinki: University of Helsinki Department of English. Available at <<http://khnt.hit.uib.no/icame/manuals/HI/INDEX.HTM>>. Accessed 17 June 2008.
- Kytö, Merja and Terry Walker (2006), *Guide to A Corpus of English Dialogues 1560–1760 (Studia Anglistica Upsaliensia 130)*. Uppsala: Acta Universitatis Upsaliensis.
- Kytö, Merja, Peter Grund and Terry Walker (2007), 'Regional variation and the language of English witness depositions 1560–1760: constructing a “linguistic” edition in electronic form', in: Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen and Jukka Tyrkkö (eds.) *Towards Multimedia in Corpus Studies (Studies in Variation, Contacts and Change in English 2)*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. <[http://www.helsinki.fi/varieng/journal/volumes/02/kyto\\_et\\_al](http://www.helsinki.fi/varieng/journal/volumes/02/kyto_et_al)>. Accessed 10 July 2008.
- Lass, Roger (2004), 'Ut custodiant litteras: Editions, Corpora and Witnesshood', in: Marina Dossena and Roger Lass (eds.) *Methods and Data in English Historical Dialectology (Linguistic Insights 16)*. Bern: Peter Lang. 21–48.
- Machan, Tim William (1994), *Textual Criticism and Middle English Texts*. Charlottesville and London: University Press of Virginia.
- 'MEG: Project Summary'. *The Middle English Grammar Corpus*. Merja Stenroos, Martti Mäkinen, Simon Horobin and Jeremy Smith (compilers). April 2008. University of Stavanger. Available at <[http://www.uis.no/research/culture/the\\_middle\\_english\\_grammar\\_project/project\\_summary](http://www.uis.no/research/culture/the_middle_english_grammar_project/project_summary)>. Accessed 19 June 2008.

- Meurman-Solin, Anneli (2007), *Manual for the Corpus of Scottish Correspondence, 1500–1730*. <<http://www.helsinki.fi/varieng/csc/manual>>. Accessed 18 August 2008.
- Nurmi, Arja (1999), 'The Corpus of Early English Correspondence Sampler (CEECS)', *ICAME Journal*, 23: 53–64.
- Rissanen, Matti (2000), 'The world of English historical corpora: From Cædmon to computer age', *Journal of English Linguistics*, 28: 7–20.
- 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). <<http://www.digitalmedievalist.org>>. Accessed 12 May 2008.
- Shillingsburg, Peter (1986), *Scholarly Editing in the Computer Age. Theory and Practice*. Athens, Georgia: The University of Georgia Press.
- Stenroos, Merja and Martti Mäkinen (2008), 'Corpus Manual, 1.0', *The Middle English Grammar Corpus*, Merja Stenroos, Martti Mäkinen, Simon Horobin and Jeremy Smith (compilers). Stavanger: University of Stavanger. <[http://www.uis.no/getfile.php/Forskning/Kultur/MEG/Corpus\\_manual\\_1.0.rtf](http://www.uis.no/getfile.php/Forskning/Kultur/MEG/Corpus_manual_1.0.rtf)>. Accessed 18 June 2008.
- Vanhoutte, Edward and Ron Van den Branden (forthcoming), 'Describing, transcribing, encoding and editing modern correspondence material: A textbase approach', *Computing the Edition. Special Issue of Literary & Linguistic Computing*. Preprint available at <<http://www.kantl.be/ctb/pub/preprint/comedvanvanfig.pdf>>. Accessed 18 August 2008.