# Research

## A European Database of Descriptors of English Electronic Texts

***Hans-Jürgen Diller**, University of Bochum, Germany*
***Hendrik De Smet**, University of Leuven, Belgium*
***Jukka Tyrkkö**, University of Helsinki, Finland*

Readers of the *Messenger* need hardly be told of the importance and impact of electronic corpora in linguistics, including historical linguistics. The very first issue (the "Zero Issue" in fact) of the *Messenger* carried an article on the compilation of the Helsinki Corpus.[4] Although a number of corpora were already in use at the time, the Helsinki Corpus was a pioneer when it came to the use of corpora in the historical study of English.

From the beginning, a key problem for corpus compilers has been the question of corpus composition. The first electronic corpus of English, the Brown Corpus (1961), used a set of textual categories which were subsequently adopted and adapted by numerous corpora compiled following the same model from LOB to FLOB, Frown and so many others.[5] Considering the rather different needs of the historical linguist, the Helsinki team led by Matti Rissanen created a set of 22 "textual parameters" which not only give textual descriptors like "Text Types" and "Prototypical Text Categories" but also includes the date, dialectal / regional provenance of the text, the author's sex and social rank, his or her relationship with the audience or readership, its relationship to foreign-language originals and to the spoken language (Kytö 1991: x-xi). Note that the criteria of 'balance and diversity',[6] usually considered essential requirements of corpus composition, are much harder to satisfy in a historical corpus. Not only is it often difficult to provide sufficient primary data as it is, particularly of the earlier periods, but text types and genres also change and develop over time. This in turn affects the way corpora can be used, for whenever a century of writing in any given genre is represented by a small handful of extracts, conclusions can only be drawn on a very tentative basis — a fact the compilers of the Helsinki Corpus openly acknowledge. As a consequence, findings based on corpus evidence run the risk of being slanted in ways that scholars may not even be aware of. We feel that this awareness ought to be fostered more.

Another problem is quantity. The Helsinki Corpus contains some 1.5 million words intended to represent the whole history of English from the earliest extant records to about

---

4 *The Helsinki Corpus of English Texts* (1991). Dept. 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). See www.helsinki.fi/ varieng/CoRD/corpora/HelsinkiCorpus/.

5 For a very full list of corpora (with descriptions) see www.helsinki.fi/varieng/CoRD/corpora/ index.html.

6 See, e.g., Biber *et al*. (1998: 246–53).

1710 (Kytö 1991: 2). Although the Helsinki Corpus has proven its worth as a valuable diagnostic corpus, the small volume of text means that rare phenomena, including many syntactic constructions, cannot be reliably studied with only that one corpus.

The problem of quantity is no longer what it was in the 1990s. Since the launching of the Helsinki Corpus the number of computer-readable historical English texts has risen to heights unforeseeable only a decade or so ago. Texts are available both from open access online repositories such as Project Gutenberg, www.gutenberg.org/wiki/ Main_Page, and the Internet Archive, www.archive.org/, and commercial repositories such as Early English Books Online (EEBO) and Literature Online (LION). The former, though freely available, are problematic for scholars due to frequently haphazard practices of editing and referencing. The provenance of a copy may be inaccurately or erroneously reported, the scanning and subsequent automated character recognition may be of a low quality and, depending on the repository, copies may even be produced by splicing text from more than one primary source. The latter sources, while providing texts of a much higher quality, can be problematic due to high pricing and copyright issues. However, even with all the remaining challenges the reality is that scholars of historical texts today have access to tens of thousands of texts right from their own computer. Things have improved in that regard.

If the problem of quantity has lost some of its seriousness, the problem of balance has arguably become worse rather than better. The freely available texts, e.g. from Gutenberg or the Internet Archive, do not show a wide range of genres and appear as an undifferentiated mass in the world-wide web (Davies 2010: 413). To create out of them a corpus that would satisfy the criteria of balance and diversity is a daunting task. Traditionally, compilers of historical corpora have spent years discussing and researching the best internal structure for a corpus in order to ensure the most accurate and reliable collection of texts. With little if any such structure available, and certainly nothing to satisfy the needs of serious scholarship, the open access repositories provide at present a tantalizing but largely unusable resource.

In our view, the solution to the problem of balance lies not primarily in the creation of a further mega-corpus but in a database of text descriptors that would guide individual researchers through the existing repositories and enable them to create their own task-specific corpora which they can balance according to their own needs and for whose composition they are individually responsible. Such corpora would benefit from an openly available, systematic, and peer-reviewed set of descriptors, all the while meeting the requirements of individual research projects. By dramatically reducing the time it takes to scour through dozens of websites and tens of thousands of texts in order to find, let us say, eighteenth century essays written by female authors, such a database would also give scholars more time to focus on research itself. With all or most of the texts used as primary data available online, transparency of research would be improved as well.

With such a database in mind, we started our European Database of Descriptors of English Electronic Texts (EUDDEET) in 2008.[7] The basic premise of the database is simple enough: to provide a stable set of textual descriptors for historical texts currently available

[7] The authors are the principal investigators of the EUDDEET project. We are grateful to Irma Taavitsainen (University of Helsinki) and Hubert Cuyckens (University of Leuven) who serve as advisors.

on open access repositories. The database allows users to identify texts that match desired descriptive criteria and provides valid download links to the respective texts in the source repositories. In due course, we hope to add more functionalities including searching within full text, zip downloads and TEI compliant XML-editions of at least some of the texts.

The first collection of texts included in EUDDEET was Hendrik De Smet's CLMETEV (Corpus of Late Modern English Texts, Extended Version, cf. De Smet 2005), a corpus of 176 texts from 1710–1920. All texts in CLMETEV were selected by De Smet from the Project Gutenberg website and other online archiving projects, making it the ideal starting point and playground for proof-of-concept. Once the descriptors were in place for CLMETEV (see below), we gradually began to add more texts using the third volume of the *New Cambridge Bibliography of English Literature* (NCBEL, Watson 1969) as our guideline.[8] The literary bias which the choice of *NCBEL* may suggest is more apparent than real, for substantial parts of *NCBEL* are devoted to journalism, the humanities and even sciences, which do not form part of literature in the more narrow sense. At present we have a database of more than 1,000 titles, certainly not bad in comparison with Helsinki's 519 text excerpts.

One of the key issues in corpus linguistic text description is the need to be able to compare results obtained from different corpora with one another. To this end, it is desirable that the compatible descriptors are used in as many corpora as possible, and that the descriptors used are defined in transparent terms. Recognising the established nature of the textual parameters used in the Helsinki Corpus, we started our work with the list of COCOA Headers ("Parameter Codes") used in Helsinki. However, over the course of applying the parameters to our collection we found that while the biographical descriptors are unproblematic, the same cannot be said for the other text descriptors. This is partly due to the greater diversity of distinctly recognizable genres in later periods of writing — bearing in mind that the Helsinki descriptors were devised for the OE to EMoDE periods — and partly due to the desire to retain a distinction between genre and topic. At the time of writing, our database follows the descriptor set given in Table 1.

Most of the descriptors in Table 1 are self-explanatory. The final descriptor, "Notes", is reserved for incidental, unsystematic information which may be found useful at a later stage. In most cases the bibliographical and biographical information has to be retrieved from authoritative sources, as the open-access repositories do not as a rule provide such information and when they do it is not always reliable.

[8] So far we have not used the corresponding volume of the third edition (Shattock 2000), since the other volumes of that edition have not been published to date; we thought it preferable at the present stage to base the entire database on one work.

| | |
|---|---|
| Title | Author's sex |
| Author's name | Author's birth |
| Prototypical Text Category (PTC) | Date first published |
| Genre | Author's death |
| Subgenre | Website |
| Verse/Prose | Notes |
| Written/Spoken | |

Table 1: Descriptors used in EUDDEET.

As noted above, the problematic categories are the ones describing the content of the text, namely PTC, Genre, and Sub-genre. Sub-genre has no counterpart in Helsinki, but we found it necessary for practical and systematic reasons which will be explained below.

The basic concept is no doubt that of *genre*, originally called "text type" in the Helsinki nomenclature. We define genres with Biber (1989: 5-6), as "text categories readily distinguished by mature speakers". This is what the psychology of categorization calls a "basic-level category". A current working list of genres includes the following:

DIARY, DRAMA, ESSAY, FAIRY TALE*, FICTION, HANDBOOK / TEXTBOOK, JOURNALISM, LECTURE, LEGEND*, LETTER, PAMPHLET, SERMON, SPEECH, TREATISE.[9] Ideally, the genre labels in our database should be intuitively clear, and this is largely true of LETTER, DIARY, SPEECH, SERMON, DRAMA (perhaps better STAGE PLAY), LECTURE, and FICTION. But some of the categories like ESSAY and TREATISE will need further consideration. We have striven to define them in more or less explicit terms in our own work, but there are plenty of texts which beg to be described as both ESSAY and TREATISE. Work is ongoing on this issue.

Proceeding from the basic level to the superordinate level, we come to what in Helsinki are called "Prototypical Text Categories" (PTC), which are intended as "larger categories which are preserved to reflect the continuity of the types of text represented throughout the history of English" (Kytö 1991: 55). There are currently six PTC categories in the database: NARRATIVE, NARRATIVE NON-IMAGINARY, EXPOSITORY, ARGUMENTATIVE, INSTRUCTIVE, and INTERACTIVE. We have found PTCs useful heuristically, particularly as far as they help reveal problems in balance. For example, even among the more than 1,000 items in the database at present, there are very few texts that would qualify as "Instruction" (whether "Religious" or "Secular"), which is an important category in Helsinki.[10] However, our present thinking is toward ultimately abandoning the PTC descriptor in our final version, mainly for reasons to do with the fact that EUDDEET descriptors apply to complete texts. In the Helsinki Corpus, which is a corpus of text excerpts, more than 40% of all text samples are left without a PTC categorization. If PTC is such a recalcitrant category in Helsinki, it is bound to be even more so in a database which will represent entire texts. For text excerpts of ca. 2,000 words a yes-or-no decision should be much easier than for texts running to perhaps half a million words (see Biber 1988 and later studies, showing that categories like NARRATIVE or EXPOSITORY are gradient rather than binary categories and may change in the course of a text).

---

9 Genres marked with an asterisk are very poorly represented at the moment and should perhaps be combined with other genres and identified as subgenres.

10 It is quite possible that the third edition of the *Cambridge Bibliography of English Literature* will produce better results in this respect (see Shattock 2001: 59).

While PTC is a superordinate category to genre, sub-genre is a subordinate one. Sub-genres are largely distinguished according to their subject-matter, hence they are to some extent identical with academic disciplines. Their importance is duly noted by the divisions of *NCBEL* (cf. Shattock 2001: *loc. cit.*). If subject-matter is used as a criterion, Biography, Autobiography, Travelogue should probably be treated as sub-genres, although they are genres in Helsinki. Sub-genres are very much an open list. Some genres will hardly lend themselves to sub-classification. Others, like Essay, Treatise and Nonfiction will cry out for it. Fiction knows some sub-genres (like epistolary novels, historical novels, children's stories) which are of great interest also to linguistic and cultural historians, but there may not be an obvious sub-genre for every fictional text. Some sub-genres will be very large (like Philosophy, History, Religion), others will remain very small.

Sub-genre does not appear in the Helsinki model,[11] but the insights it offers justify its place in the list. On the one hand, sub-genre information may draw attention to biases or historical imbalances within the genre categories. On the other, it is an important place for local discoveries. Whether a word or an expression has become part of a specialized register, for instance, can be decided only if those registers have a place in the database. To give only one illustrative example: it is certainly worth knowing whether a word or a construction is peculiar to, or preferred by, let's say philosophical or historical texts.

Returning to the crucial question of balance, as a database to be used for building corpora rather than a corpus, EUDDEET need not satisfy the same criteria as a balanced corpus, but clearly a database is only of use if the texts described in it provide enough diversity and quantity to allow users to build a wide variety of different corpora.

Even though our system of categories is still in need of refinement, some general statements are possible at this stage already. The first and foremost concerns the apparent imbalances in the collection of texts. These are at this time principally a reflection of the biases in *NCBEL 3*, and will in due course be remedied as other sources of descriptive information are included. And, bearing in mind that EUDDEET only includes texts available from open access repositories, the current composition of the database also reflects to some extent the decision made by editors of and contributors to such repositories. With these reservations we feel justified in making the following claims about our categorizations so far: a fair number of genres, including DRAMA, FICTION, NONFICTION (NON-IMAGINATIVE NARRATION), and LETTER are very well represented: each of them occurs more than 50 times in the 19th century part of our database alone, while others like SERMON, SPEECH, and DIARY are severely underrepresented. The database is thus not yet as balanced as we would wish, but one of the benefits of the database format is that we are aware of the imbalance, which will induce us to look elsewhere for texts representing these genres. Some imbalances, such as the severe underrepresentation of female authors,[12] will have to be repaired by material from other sources, e.g., from websites specializing in women writers.

---

11 While subgenre is not a formal category in the Helsinki Corpus, Nevalainen and Raumolin-Brunberg (1993) discuss the concept as it pertains to the Early Modern part of the corpus.

12 Female authors currently account for approximately 20 per cent of fiction texts and 5 per cent of non-fiction texts in EUDDEET.

So far, the database and the descriptor set have been tested explicitly in one study of semiotic sign terms presented by Tyrkkö at the 16th ICEHL conference held August 23–27, 2010, in Pécs, Hungary.[13] The study, a continuation of earlier work on sign terms in Early Modern medical writing (Tyrkkö 2006a, 2006b, forthcoming), looked at the frequencies and semantic differences of terms such as *token*, *sign*, *symptom*, *mark*, etc. in texts from three Prototypical Text Categories: NARRATIVE IMAGINARY, NARRATIVE NON-IMAGINARY, and EXPOSITORY. The pilot corpus derived from EUDDEET consisted of 97 texts and slightly over eight million words.

The results showed that considerable diversity is to be found in the way sign terms are used in different styles of writing, and also that to uncover patterns of use it will be necessary to make use of the more fine-grained levels of text description, that is, genre and sub-genre. For example, although the seemingly medical term *symptom* was found to occur at a similar frequency in all PTCs, the uses differed considerably with the medical sense predominating in the NARRATIVE texts, while EXPOSITORY texts showed partiality to a more metaphorical sense. A closer look explained the difference: while NARRATIVE texts often reported on the medical condition of a character or real person, the latter particularly in biographies included in the NARRATIVE NON-IMAGINARY section, the EXPOSITORY texts generally discussed more abstract issues and *symptom* was the term of choice for a predictive sign, usually of an undesirable event or state of being. It is worth mentioning that medical texts were not included in the pilot study.

In terms of testing the database, the exercise proved valuable in pointing out certain needs in terms of functionality and the composition of the descriptor set. To give an example, the low quality of OCR scanning in many texts became painfully apparent and prompted us to start taking steps to ensure a quality rating system is in place by the time the database is released to public use. The fact that the PTC category was used as the primary compositional criterion with some success need not be interpreted as an endorsement for the descriptor, as the same results can be obtained when the PTCs are replaced by conjunctions of the appropriate genres.

The above must have shown that the projected database is still very much a work in progress. We present it here with three hopes. First, we hope to stimulate discussion among ESSE members about questions which concern us all. Text categorization is an important part of our daily work, no matter which field of English Studies we are engaged in cultivating, and feedback is welcome. The second hope is to arouse an interest in our database, part of which (perhaps the 18th and 19th century) we expect to make available in 2011. We also hope to provoke discussion among users about better categorization and better representation. There will be gaps, and there will be poor texts, for it must be remembered that the growth in the quantity of online texts which we have seen in recent years has in some respects led to a loss in quality. The results of the Optical Character Recognition methods used in the production of many digitized texts are uneven, and though we will try to pick the best, we will not always succeed. Feedback from users will be

[13] It goes without saying that where the current descriptor set is similar to the one used in the Helsinki Corpus, a staggering amount of research has successfully relied on it over the past 20 years.

essential. What we hope to gain is a more articulate and more explicit understanding of the categories which we employ in our daily work, as linguists as well as students of literature and culture. A database like the one proposed here could become an important instrument in creating, reflecting and articulating the terminological consensus which an academic discipline needs for its own cohesion.

**References**

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge UP.

Biber, Douglas. 1989. A Typology of English Texts. *Linguistics* 27: 3–43.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge UP.

Davies, Mark. 2010. Using large and diverse online corpora. *International Journal of Corpus Linguistics* 15.3: 412-18.

De Smet, Hendrik. 2005. A corpus of Late Modern English texts. *ICAME Journal* 29: 69–82.

Kytö, Merja. 1991. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts.* Helsinki: Department of English, University of Helsinki.

Kytö, Merja, & Matti Rissanen. 1990. Empirical Evidence for the Study of the Structure of English: Helsinki Corpus of English Texts: Diachronic and Dialectal. *The European English Messenger*, Zero Issue, Autumn 1990: 22–26.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 1993. Early Modern British English. In Rissanen Matti, Merja Kytö and Minna Palander-Collin (eds.) *Early English in the computer age: explorations through the Helsinki corpus.* Mouton de Gruyter. 55-73.

Shattock, Joanne (ed.). [3]2000. *The Cambridge Bibliography of English Literature*. Vol. 4: *1800-1900*. Cambridge: Cambridge UP.

Shattock, Joanne. 2001. Revising *The Cambridge Bibliography of English Literature.* Bibliographical Research in an Electronic Age. *The European English Messenger* 10/2: 56–61.

Tyrkkö, Jukka. 2006a. From *tokens* to *symptoms*: 300 years of developing discourse on medical diagnosis in English medical writing. Dossena Marina and Irma Taavitsainen (eds.) *Diachronic Perspectives on Domain-Specific English.* Bern: Peter Lang. 229-255.

Tyrkkö, Jukka. 2006b. Tokens, signs, and symptoms: Signifier terms in medical texts from 1375 to 1725. McConchie R.W., Heli Tissari, Olga Timofeeva and Tanja Säily (eds.) *Selected Proceedings of the 2005 Symposium on New Approaches in English Historical Lexis (HEL-LEX).* Somerville, MA: Cascadilla Press. 155-165.

Tyrkkö, Jukka. Forthcoming. Sign terms in specific medical genres in early modern medical texts. Taavitsainen Irma and Päivi Pahta (eds.) *Early Modern English Medical Texts: Corpus Description and Studies.* Amsterdam and Philadelphia: John Benjamins.

Watson, George (ed.). 1969. *The New Cambridge Bibliography of English Literature*, Vol. 3: 1800-1900. Cambridge: Cambridge UP.