Building a bridge between the present and the past: A corpus of 19th-century English

Merja Kytö
Uppsala University

Juhani Rudanko
University of Tampere

Erik Smitterberg
Uppsala University

1 The need for a corpus of 19th-century English

The 19th century was an age of exploration and new discoveries; yet the English language in that period remains largely an unexplored territory. Happily, over the past few years we have witnessed an increasing interest in late Modern English.[1] We have also seen a growing number of corpora being compiled and becoming available for various periods of the history of English. However, we still need access to corpora covering the 19th century.[2] There are several ways in which such corpora would be useful. First of all, they could provide the data necessary for carrying out follow-up studies to research done on early Modern English. This would be valuable for a number of syntactic studies, for instance on the distribution of modal auxiliaries and relative pronouns, various ways of expressing future time, and the development of the progressive form. In order to serve this function, however, the corpus would have to be compiled with a cross-genre perspective in mind, providing the researcher with texts stratified according to extralinguistic criteria roughly comparable to, for instance, the Helsinki Corpus of English Texts.

Secondly, there is a scarcity of corpora covering the period immediately before Present-Day English.[3] A corpus of 19th-century English would thus provide researchers with the possibility of extending studies both of short-term diachronic change and of trends in Present-Day English backwards in time. Studies

based on 19th-century corpora would thus have the advantage of a double perspective, that is, both forwards and backwards in diachrony. Here also there is a need for a cross-genre perspective to match, for instance, the LOB and Brown corpora.

Finally, many transplanted varieties of English originate in the late Modern English period; and interest in the origin and/or status of such transplanted varieties has increased of late (see eg Burchfield 1994, Hundt 1999). Studies of transplanted varieties of English would benefit from being able to draw on data concerning the status of British English in the 19th century for comparison.

It was mainly to provide for the needs described above that a project aiming at a stratified computerized corpus of 19th-century English (CONCE) was launched at the University of Tampere in the mid-1990s. The project is currently underway at the department of English Philology of that university and at the Department of English of Uppsala University. The CONCE corpus comprises 19th-century English English texts (with occasional dialectal features); the wordcount is at present a little above one million words.

2 The corpus of 19th-century English

2.1 Period division

The texts that constitute CONCE are divided into three subperiods: period 1 includes texts published between 1800 and 1830, period 2 between 1850 and 1870, and period 3 between 1870 and 1900; the subperiods thus represent, broadly speaking, the beginning, middle, and end of the 19th century. The delimitation of the subperiods was partly governed by the availability of suitable texts (an effort was made to use the earliest possible editions of printed works); however, the periods also correspond roughly to extralinguistic events of importance in Great Britain, some of which are given in Table 1.

As can be seen from Table 1, period 1 corresponds roughly to the period before the great reforms. Period 2 begins when some reforms (eg the First Reform Bill and the penny post) have had time to take effect, and coincides with a number of other important reforms (eg the Repeal of the Stamp Act and the Second Reform Bill). Period 3, finally, is more stable, with only a few reforms, and constitutes the 19th-century part of '[l]ate Victorian imperialism and the last phase of global "stability"' (Görlach 1999: 6).

Table 1: Some sociocultural events of the 19th century (taken from Harvie 1992, Matthew 1992 and Görlach 1999)

Year	**Event(s)**
1824	The repeal of the Combination Acts
1832	The First Reform Bill
1833	The first important Factory Act restricting child labour
1834	Slavery abolished; the Poor Law Amendment Act
1840	The penny post introduced
1855	The final repeal of the Stamp Act of 1712
1867	The Second Reform Bill
1870	The Elementary Education Act
1884–5	The Third Reform Bill

It is of course an open question to what extent large-scale extralinguistic events such as the above influence linguistic change that can be detected in a corpus. However, it is possible that for instance the Reform Bills, which extended the franchise, had an effect on the political language of the time; moreover, the type of printed matter produced might have been affected by the increase in literacy and consequent diversification of the readership which was a direct or indirect effect of such reforms as the penny post, the final repeal of the Stamp Act, and the Elementary Education Act.

There are also stylistic reasons for the period delimitation chosen. As Görlach (1999: 158f) states, '[e]arly 19th-century writers largely built on 18th-century foundations, in literary and expository texts, and this continuity makes many pre-1830 texts look quite "classical"'. Several style-related research questions can thus be based on contrasting the beginning of the 19th century with the rest of the century.

2.2 Genres

CONCE contains seven genres: Debates, Trials, Drama, Fiction, Letters, History and Science. An overview of the contents of these genres is given in Table 2:

Table 2: Description of the genres in CONCE

Genre	**Characteristics**
Debates	Recorded debates from the Houses of Parliament
Trials	Trial proceedings (in dialogue format)
Drama	Prose comedies or farces (in domestic etc settings)
Fiction	Novels
Letters	Personal letters (between relatives or close friends)
History	Historical monographs
Science	Monographs pertaining to the natural or social sciences

The choice of texts included makes it possible to apply a cross-genre perspective to studies of 19th-century English, as the genres represented differ with respect to a number of extralinguistic parameters, such as:

1. Medium: written to be read[4] (Fiction, Letters, History, Science); written to be spoken (Drama); speech taken down[5] (Debates, Trials).
2. Publicity: produced for publication (Drama, Fiction, History, Science); produced with speakers aware of possible/certain publication (Debates, Trials); not produced for publication (Letters).
3. Narrativity: containing passages where narrative elements dominate (Fiction), where expository elements dominate (Science) and where a mixture of narrative and other elements can be found (Debates, Trials, Drama, Letters, History).

However, the number and length of texts vary with genre. For each of the three subperiods, three texts of c 10,000 words each for Drama, Fiction, History, and Science were sampled; for Debates, one text[6] of c 20,000 words; for Letters, ten texts – five by women and five by men letter-writers[7] – of c 10,000 words; and for Trials three or four texts, together comprising c 60,000 words.

The texts in CONCE have been coded using text-level and reference codes based on those applied to the Helsinki Corpus (for which see Kytö 1996). If the same principles are followed when determining the wordcount as were applied to the Helsinki Corpus (see Kytö 1996: 168), CONCE contains 1,030,409 words.[8] The period/genre breakdown is given in Table 3:

Table 3: Wordcounts for period, genre, and period/genre subsamples in CONCE and for the whole corpus, using the same principles as those applied to the Helsinki Corpus

Period	Debates	Trials	Drama	Fiction	Letters	History	Science	Totals
1	20,123	69,908	36,524	42,032	122,669	30,904	38,037	360,197
2	20,036	63,023	38,929	39,045	133,308	30,504	31,679	356,524
3	20,286	71,713	37,896	30,113	92,513	30,564	30,603	313,688
TOTALS	60,445	204,644	113,349	111,190	348,490	91,972	100,319	1,030,409

However, the coding scheme applied to CONCE is more thorough than was the case for the Helsinki Corpus. Text-level codes have been applied to make it possible to exclude eg stage directions in Drama, indications of speakers in Drama, Debates and Trials, and address and date information in Letters. If this coding scheme is used for the wordcount, the figures decrease especially in speech-based genres. The period/genre breakdown is given in Table 4.

Table 4: Wordcounts for period, genre, and period/genre subsamples in CONCE and for the whole corpus, excluding the words within reference codes and text-level codes

Period	Debates	Trials	Drama	Fiction	Letters	History	Science	Totals
1	19,908	62,360	31,311	42,032	121,624	30,904	38,037	346,176
2	19,385	60,570	29,543	39,045	131,116	30,504	31,679	341,842
3	19,947	67,588	29,090	30,113	90,891	30,564	30,603	298,796
TOTALS	59,240	190,518	89,944	111,190	343,631	91,972	100,319	986,814

2.3 Status

Within the CONCE project, the collection of texts has now been frozen, but the second round of proof-reading has not yet been carried out; nor have copyright issues been tackled so far. For these reasons, CONCE is still an in-house corpus at Uppsala University and the University of Tampere.

3 Special research possibilities

Although CONCE has been compiled to give a good general coverage of 19th-century English, there are also special features of the corpus which make it particularly suitable for addressing certain more specific research questions. Examples of these are presented in this section.

3.1 In-depth study of the Letters genre

As the Letters genre constitutes slightly more than a third of CONCE, it is well suited for in-depth, single-genre studies of individual linguistic features. Letters appropriately represent language use in general, as they in many ways reflect a middle-ground between written and spoken characteristics, being a written genre often influenced by spoken informal features. Moreover, letters are pointed out by Görlach (1999: 149) as being of particular interest 'since they reflect the social and functional relations between sender and addressee to a very high degree – only spoken texts can equal this range'.

Since the Letters genre has been compiled with the gender parameter in mind,[9] it is possible to study at least moderately frequent linguistic features within the framework of gender studies using this genre (for wordcounts in Letters, see Table 5). In fact, clear gender-related differences can be found even for low-frequency features such as the progressive form (see Arnaud 1998: 139f, who found clear differences between women and men as regards the use of the progressive form in his corpus consisting of private letters).

Table 5: Wordcounts for the letters by women and men writers in CONCE, excluding the words within reference codes and text-level codes

Period	Women	Men
1	69,271	52,353
2	62,340	68,776
3	50,154	40,737

3.2 Speech-related genres

CONCE contains four genres which can be said to be partly or wholly speech-related: Fiction (dialogue written to represent speech), Drama (dialogue written to be spoken), Debates (speech taken down as indirect and direct speech) and Trials (speech taken down as direct speech). Of these, the latter three have been

coded so that it is possible to include only speech-based material as such in the wordcounts and searches for data.[10]

Of course, none of these genres can fully make up for the lack of access to the natural speech of the period. The Debates genre, for instance, consists of formal political speech, and Görlach (1999: 149), commenting on speeches in general as a genre, states that '[a]lthough the heyday of rhetoric in English "liberal education" was the Renaissance … there was enough of the tradition left in the 19th century to make some texts markedly more "artificial" than modern ones' (see also Note 5 above). However, it is to be hoped that by extrapolating from these genres, which all have some bearing on 19th-century speech, researchers will be able to obtain a better picture of the spoken language of the Victorian period than has hitherto been the case. Moreover, although the texts have not been stratified according to gender in the same way as those of the Letters genre have, an effort has still been made to ensure that Fiction, Drama and Trials contain women as well as men authors and speakers. With some additional work on the part of the researcher, it is thus possible to study differences between constructed or authentic dialogue produced by women and men.

3.3 Academic language

When investigating the language of the LOB and LLC corpora, Biber (1988: 171) found that academic prose contains more genre-internal linguistic variation than, for instance, personal letters. He also states that academic prose texts comprise 'several well-defined sub-genres', and that 'the variation within the genre is due in part to variation among the sub-genres'. This makes it an interesting question whether such diversity is a recent phenomenon, or whether similar variation can be attested for the 19th century. Moreover, Görlach (1999: 150) claims that '[a] few expository and literary text types which originated or were redefined in the 19th century deserve closer inspection. They include scientific style which changed from somewhat personal accounts to impersonal, objective description'. The above statements make 19th-century academic writing a very interesting topic for studies in diachronic variation, both forwards and backwards in time.

CONCE includes two genres that belong to the umbrella genre 'academic writing': History and Science. Of these, Science in turn includes the two subgenres 'natural science' and 'social science'; natural science texts make up the bulk of the genre, however. CONCE can thus be used to study variation within the field of academic writing in at least three ways:

1. Academic writing from before 1800 can be compared to corresponding genres in CONCE to see how the developments mentioned by Görlach affected especially perhaps the Science genre; comparisons with present-day conventions of academic writing are also possible.
2. Biber's (1988) factor analysis, or a modified version of the same, can be applied to CONCE to see how big the genre diversity was within the umbrella genre 'academic writing' in the 19th century. Moreover, Biber and Finegan's (1997) results for the Science texts in ARCHER can be tested against the CONCE data.
3. The different subgenres of academic writing included in CONCE can be compared and contrasted with each other to see whether there are as clear differences between them, and whether these differences, if any, increase or decrease in diachrony.

3.4 Short-term linguistic change

As the genre/period subsamples distinguished in CONCE (1800–30, 1850–70 and 1870–1900) contain roughly the same number of texts of roughly the same length, it is possible to conflate the subsamples and treat the 19th century as a whole, which may doubtless be useful when the researcher wishes to compare the language of the 19th century to previous or later periods. However, it is also possible to keep the subsamples apart and focus on linguistic development within the 19th century. It is to be hoped that this possibility will make it easier for scholars to locate possible causes of linguistic change. In this respect, CONCE provides a rough 19th-century equivalent of the LOB, FLOB, Brown and Frown corpora.[11] Studies based on these corpora have shown that a difference of c 30 years is enough to study linguistic change. CONCE thus ties in with the recent scholarly interest in short-term change in diachrony.

4 Pilot studies

So far, a number of pilot studies have been carried out or are underway on CONCE, either on the corpus as a whole or on parts of the corpus, sometimes in comparison with other corpora. Glimpses at the results obtained are given below.

Rudanko (1998, Chapter 3) presents work done on complementation in 19th- and 20th-century British English. When that study was written, Fiction, History, and Science were almost in their final form, and these genres were drawn on for data for the 19th century, with the data for the 20th century coming from the G and K parts of the LOB corpus.

The investigation focuses on the basic subject control pattern of the *to* infinitival type, as in *John decided to leave*, and examines the types of matrix verbs that select this pattern in 19th- and 20th-century English. The study makes use of the the kinds of semantic concepts introduced in Rudanko (1989), including those of desideration, intention, decision, and endeavour or effort, and suggests that these notions are relevant to the analysis of matrix verbs in both centuries. It is also argued that these concepts lend themselves to the investigation of diachronic change affecting this part of English grammar. Such change is seen to be of different types. For instance, a major class of matrix verbs selecting the pattern in both the 19th and the 20th centuries is that of verbs of decision, but the internal makeup of the class has undergone considerable change. Thus in 19th-century English, the verbs *determine* and *resolve* were very common with *to* infinitives, whereas the verb *decide* was quite rare. However, in the LOB material, these roles are reversed, and *decide* has become a high-frequency item in the pattern, whereas *determine* and *resolve* have become rare by comparison.

Overall, the investigation, with its emphasis on the use of corpus data, sheds light on change in the semantic and syntactic nature of matrix verbs selecting the *to* infinitival pattern during the last two centuries.

Smitterberg (forthcoming) investigates the frequency and clausal distribution of the progressive form in Drama, Fiction, History, Letters and Science. The study shows that there are considerable and consistent genre differences, which seem to increase over time, and where speech-based and/or informal genres appear to favour the use of the progressive. As concerns clausal distribution, genres which favour the progressive form also favour its occurrence in main clauses. When all genres are taken together, there is a general trend across the century towards increasing frequency of the progressive and increasing occurrence of the progressive in main clauses.

Smitterberg, Reich and Hahn (2000, in this issue of the *ICAME Journal*) look into the frequency and adverbial modification of the present progressive from a cross-genre and cross-corpus perspective: Debates and Science in CONCE are compared with genres consisting of Present-Day English political and academic language in the English–German Translation Corpus. Although frequencies are too low for results to be conclusive, and differences between the corpora used may also affect the results, the study indicates that the progressive has increased more in political than in academic language over the past two centuries. Moreover, a look at the modification of the present progressive form by temporal adverbials indicates that, at least in academic language, such modification may have decreased in diachrony.

Geisler (in preparation) focuses on register variation in 19th-century English, as based on the version of CONCE tagged using the EngCG–2 tagger. The investigation is based on an analysis of so-called dimension scores, using Biber's dimensions of linguistic variation (Biber 1988). By computing dimension scores from standardized frequencies of a number of salient grammatical features, it is possible to compare the distribution of dimension scores both between registers and across subperiods. In addition, the study also compares the CONCE results with previous diachronic register analyses as reported in Biber (1995), and in Biber and Finegan (1989, 1997). Conclusions are drawn on the suggested 'drift' developments in oral and literate genres.

In sum, the above studies give an idea of the potential of the CONCE corpus as a tool for linguistic investigations. In their further work, the members of the project team and their collaborators will address various research topics, with an overall aim to account for the 'essentials' of English in the 19th century, as against the past and subsequent development of the language.

Notes

1. See, eg, Bailey (1996), Romaine (1998) and Görlach (1999).
2. The existing ARCHER corpus, which covers the period 1650–1990 is of course very important in this respect (see eg Biber and Finegan 1997). ARCHER provides the researcher with the possibility of carrying out cross-genre studies, and also of comparing British and American English (though American English texts have so far been sampled only for every second subperiod, and the Science genre includes only British English texts). The speech-related registers are represented by drama, sermons and quoted speech in fiction ('fictional conversation'); other registers include fiction (other than quoted speech), letters, journals or diaries, legal opinion, news, medical texts and scientific reports from the *Philosophical Transactions of the Royal Society*. For each 50-year period ARCHER includes about 20,000 words per register (Biber, Finegan and Atkinson 1994: 3–4).
3. In the family of single-genre corpora, the corpus of late Modern English prose compiled by David Denison offers access to some 100,000 words sampled from 1860 to 1919 (Denison 1994).
4. In the 19th century it was also quite a common practice to read both letters and fiction texts out loud for eg the members of the reader's family. However, it is probable that the writers of novels and letters mainly meant their texts to be read in silence.

5. This classification, however, must not be taken to mean that the texts in these two genres constitute samples of actual speech. First, parts of Debates are rendered as indirect rather than direct speech, thus increasing narratorial presence. Secondly, the researcher should always keep in mind possible effects of the editorial process 19th-century texts went through when being transferred into written form.
6. The notion of 'text' is complicated in connection with the Debates genre. It could be argued that each speech constitutes a text, in which case the decision was rather to sample several complete texts (speeches) until the word-count reached 20,000. The same is true to some extent of Letters and Trials.
7. Period 3 of the Letters genre contains only four texts by women letter-writers, and consequently nine texts altogether (a misleading period code made it necessary to remove one text from the collection).
8. As the second round of proof-reading and possible re-coding has not yet been carried out, all wordcounts should be considered approximate.
9. Efforts have been made to include female voices in all genres where this has been possible, eg women witnesses in Trials, women characters in Drama, and women writers in Fiction. As concerns Letters, these efforts have resulted in a stratified subcorpus which can, as such, be used for gender studies; every subperiod also contains a Fiction text by a woman writer, but the total number of texts for this genre is probably too low to rule out idiolectal influence.
10. However, for the sections of Debates that consist of indirect speech, information about speakers has not been coded, as such information is part of the text proper from a syntactical point of view.
11. It should be noted, however, that the texts that constitute CONCE were produced over a period of years rather than during one single year, as is the case for LOB, FLOB, Brown and Frown; this difference may make subperiod comparisons less exact; Smitterberg (forthcoming) nevertheless found quite stable differences in the frequency of the progressive form across the subperiods.

References

Arnaud, René. 1998. The development of the progressive in 19th century English: A quantitative survey. *Language Variation and Change* 10, 123–152.

Bailey, Richard W. 1996. *Nineteenth-century English*. Ann Arbor: The University of Michigan Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65: 487–517.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie and P. Schneider (eds) *Creating and using English language corpora,* 1–13. Amsterdam–Atlanta, GA: Rodopi.

Biber, Douglas and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds) *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253–275. Mémoires de la Société Néophilologique de Helsinki, Tome LII. Helsinki: Société Néophilologique.

Burchfield, Robert W. (ed). 1994. *The Cambridge history of the English language V: English in Britain and overseas*. Cambridge: Cambridge University Press.

CONCE = A Corpus of Nineteenth-Century English, being compiled by Merja Kytö (Uppsala University) and Juhani Rudanko (University of Tampere).

Denison, David. 1994. A corpus of late Modern English prose. In M. Kytö, M. Rissanen and S. Wright (eds) *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March 1993,* 7–16. Amsterdam and Atlanta, GA: Rodopi.

Geisler, Christer. In preparation. Investigating register variation in nineteenth-century English: A multi-dimensional comparison. A paper to be presented at The Second North American Symposium on Corpus Linguistics and Language Teaching, Northern Arizona University (Flagstaff AZ), 31 March–2 April, 2000.

Geisler, Christer, Merja Kytö and Erik Smitterberg. Forthcoming. Linguistic features and dimensions 1800–1900: Ongoing research. Paper to be presented at the 21st ICAME Conference, Sydney, Australia, 21–25 April, 2000.

Görlach, Manfred. 1999. *English in nineteenth-century England: An introduction*. Cambridge: Cambridge University Press.

Harvie, Christopher. 1992. Revolution and the rule of law (1789–1851). In K. O. Morgan (ed) *The Oxford illustrated history of Britain*, 419–462. London, New York, Sydney, and Toronto: BCA.

Hundt, Marianne. 1999. *New Zealand English grammar: Fact or fiction? A corpus-based study in morphosyntactic variation*. Varieties of English Around the World 23. Amsterdam and Philadelphia: John Benjamins.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. 3rd ed. Helsinki: Department of English, University of Helsinki.

Matthew, H. C. G. 1992. The liberal age (1851–1914). In K. O. Morgan (ed) *The Oxford illustrated history of Britain*, 463–522. London, New York, Sydney, and Toronto: BCA.

Romaine, Suzanne (ed). 1998. *The Cambridge history of the English language IV: 1776–1997*. Cambridge: Cambridge University Press.

Rudanko, Juhani. 1989. *Complementation and case grammar*. Albany, New York: State University of New York Press.

Rudanko, Juhani. 1998. *Change and continuity in the English language*. Lanham, New York: University Press of America.

Smitterberg, Erik. Forthcoming. The progressive form and genre variation during the nineteenth century. In R. Bermúdez-Otero, D. Denison, R. M. Hogg and C. B. McCully (eds) *Generative theory and corpus studies: A dialogue from 10ICEHL*. Topics in English Linguistics 33. Berlin and New York: Mouton de Gruyter.

Smitterberg, Erik, Sabine Reich and Angela Hahn. 2000. The present progressive in political and academic language in the 19th and 20th centuries: A corpus-based investigation. *ICAME Journal* 24: 99–118.