A Corpus of English Dialogues 1560–1760 (CED)

The CED was compiled as a tool for the study of the language of the Early Modern period; the focus was placed on dialogues because interactive face-to-face communication is known to be an important factor in language change. The corpus was designed to offer easy access to a substantial quantity of data for variationist studies and research into historical pragmatics, as well as the study of speech presentation: it was compiled with particular variables in mind, such as text type, time, gender, and social rank. As the CED focuses on spoken interaction in the past, it facilitates the study of topics such as politeness phenomena, and conversational structure. The CED also includes various modes of speech presentation, e.g.**,** direct and indirect speech, making the material of special value to those investigating how speech is presented in writing.

The CED is part of the research project "Exploring spoken interaction of the Early Modern English period (1560–1760)". The CED is available in two formats: plain text and XML. Both the plain text and the XML versions comprise 177 individual text files.

Released in Spring 2006, ***A Corpus of English Dialogues 1560-1760 (CED)*** is a 1.2-million-word computerized corpus of Early Modern English speech-related texts. The CED is part of the research project “Exploring spoken interaction of the Early Modern English period (1560-1760)", and was compiled by Merja Kytö and Jonathan Culpeper, in collaboration with Terry Walker and Dawn Archer, at Uppsala and Lancaster Universities. In the following, we explain the background and structure of the corpus. We comment on the text types, sampling criteria, and coding. We also give word counts for the entire corpus and the stretches coded for direct speech.

The CED was compiled as a tool for the study of the language of the Early Modern period; the focus was placed on dialogues because interactive face-to-face communication is known to be an important factor in language change. The corpus was designed to offer easy access to a substantial quantity of data for variationist studies and research into historical pragmatics, as well as the study of speech presentation: it was compiled with particular variables in mind, such as text type, time, gender, and social rank. As the CED focuses on spoken interaction in the past, it facilitates the study of topics such as politeness phenomena, and conversational structure. The CED also includes various modes of speech presentation, e.g., direct and indirect speech, making the material of special value to those investigating how speech is presented in writing.

There are 177 text files in the CED, yielding a total of 1,183,690 words. The CED contains texts representative of five text types (plus a mixed bag of dialogues labeled ‘Miscellaneous’), which divide into two categories: these are ‘authentic dialogue’, which is written records of real speech events (Trial Proceedings and Witness Depositions), and ‘constructed dialogue’, in which the dialogue is constructed by an author (Drama Comedy, Didactic Works, and Prose Fiction). Furthermore, the text types may consist of dialogue in which the intervention of the narrator is minimal, limited to identifying the speaker or

marking scene changes and the like, as in Drama Comedy, Didactic Works, and Trial Proceedings, whereas other text types contain considerable intervention by the scribe or narrator, with dialogue embedded in a third person narrative, as in Witness Depositions and Prose Fiction

The text types can be briefly described as follows. Trial Proceedings are reports of the proceedings in court, typically recorded by an official scribe (records written by those involved in the trial, such as the defendant, were excluded from the CED). The dialogue is recorded as direct speech, generally in the form of questions and answers. Witness Depositions are written records of the oral testimony of witnesses, usually given before the trial proceedings themselves, which are rendered by a scribe as a third-person narrative, with legal formulae inserted. Occasionally, dialogue from earlier speech events cited by a witness is rendered as direct speech by the scribe. Drama Comedy contains dialogue in the form of direct speech, invented by an author. The text-type Didactic Works also consist of texts constructed by an author with the dialogue presented as direct speech. These are handbooks and instructional treatises, typically containing dialogues between 'instructor' and 'instructee'. Language teaching handbooks are a small group of texts set apart from the other Didactic Works (which are hence labeled 'Other' i.e. 'other than language teaching'): the language of the dialogue can be contrived for didactic purposes and is likely to be influenced by both the target language and the author's native language. Prose Fiction texts include fictional, constructed dialogue, but unlike Drama Comedy the dialogue can be presented in the form of direct or indirect speech, and is surrounded by narration by the 'storyteller'. 'Miscellaneous' is not a text type at all, but a collection of dialogues presented as direct speech which could not be classified as belonging to any of the above text types.

Table 1 illustrates the overall structure of the CED outlined above, and also gives the overall word counts for each text type and the Miscellaneous texts. (These counts were obtained using the Hcount computer program, excluding foreign language, editorial comments, and text added by the corpus compilers.)

Table 1: Overall structure of the CED and word counts for each text type (and Miscellaneous texts)

	Authentic dialogue	**Constructed dialogue**
Minimum narratorial intervention	Trial Proceedings 285,660 words	Drama Comedy 238,590 words
		Didactic Works A. Other 162,250 words B. Language Teaching 74,390 words
		Miscellaneous 25,970 words
Considerable narratorial intervention	Witness Depositions 172,940 words	Prose Fiction 223,890 words
Total word count	458,600	725,090

In the CED, the 200 years 1560-1760 is divided into five 40-year periods, as shown in Table 2. This table also gives the word counts for each period for all text types (and Miscellaneous texts) taken together.

Table 2: The periodization of the CED and the period word counts (for all five text types plus Miscellaneous texts)

Period	Period totals
1 1560-1599	200,150
2 1600-1639	204,470
3 1640-1679	259,240
4 1680-1719	297,090
5 1720-1760	222,740
Total	1,183,690

Where possible, we sampled one or more extracts amounting to around 10,000 words from each text. However, shorter texts that otherwise fulfilled our selection criteria were also used. The main criteria for selection were that the texts should:

- belong to one of the text types described above
- include speech presentation, preferably in the form of direct speech
- preferably include speakers of both sexes
- preferably include speakers representative of a range of social ranks
- represent the language of the period 1560-1760
- preferably be the earliest extant printed version.

Some Trial and Deposition texts had no printing contemporaneous with the speech event, and in this case, later text editions were used, providing these could be verified against a manuscript record (time and funding considerations prevented the use of manuscript records as CED source texts).

Coding in the CED has been added to offer computerized texts which remain as close as possible to the original versions while facilitating computer searches. Font changes, foreign language, headings, editorial comments, and text added by the compilers (comments on source text peculiarities, etc.) have been marked off from the running text by coding. As the CED is primarily designed for the study of dialogue, the compilers have taken two additional measures: long narrative passages have been omitted and replaced by a summary (coded as

compilers' comments), and direct speech has been distinguished from the rest of the running text. The latter was done by marking off text other than direct speech, although it was not always unproblematic to mark e.g., where indirect speech ended and direct speech began. Based on our interpretation, in Table 3 we give the overall word counts for direct speech in the five 40-year periods.

Table 3: Period word counts for direct speech (for all five text types plus Miscellaneous texts)

Period	Period totals
1 1560-1599	140,410
2 1600-1639	145,880
3 1640-1679	192,150
4 1680-1719	237,030
5 1720-1760	178,630
Total	894,100

The CED is intended for non-commercial research or teaching purposes only. It has been deposited at the Oxford Text Archive (see http://ota.ahds.ac.uk/) and will be available on the forthcoming ICAME CD-ROM. The CED exists in two formats: plain text and XML. Both the plain text and the XML versions comprise 177 individual text files.

References

A Corpus of English Dialogues 1560-1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).

Culpeper, Jonathan and Merja Kytö. 2000. Data in historical pragmatics: Spoken interaction (re)cast as writing. Journal of Historical Pragmatics 1 (2): 175-199.

Culpeper, Jonathan and Merja Kytö. 2010. Early Modern English Dialogues: Spoken Interaction as Writing. Cambridge: Cambridge University Press.

Kytö, Merja and Terry Walker. 2006. Guide to A Corpus of English Dialogues 1560-1760 (Studia Anglistica Upsaliensia 130). Uppsala: Acta Universitatis Upsaliensis.