Medicor: A corpus of contemporary American medical texts

Minna Vihla
University of Helsinki

1 Introduction

Medical texts are the medium in which new medical hypotheses are formulated. They are also a means of distributing medical knowledge to the general public, and they can be a form of linguistic manipulation which tries to influence the reader's future action. These aspects make medical discourse an interesting research area for a linguist. Using corpus linguistic methods, it is possible to study quantitatively the interplay between form and function in medical texts.

This paper introduces a new computer corpus which will enable quantitative study of medical discourse. The new corpus, *Medicor*, contains contemporary American medical texts, and its size is 397,311 words. It is being compiled at the University of Helsinki by Minna Vihla.[1] *Medicor* represents different types of medical writing, both professional and popular: samples taken from medical textbooks, professional handbook samples, research and editorial articles published in professional medical journals, samples from a popular medical guidebook, and newspaper/magazine articles intended for the general public. In what follows, section 2 briefly presents the background of the corpus and its place in the corpus linguistic setting, whereas later sections describe the corpus itself.

2 Why a new corpus?

Can a special computer corpus of medical texts be justified? I would say it can. Medical writing is, on the one hand, a form of scientific discourse. On the other, it includes texts illustrating how medical knowledge is distributed between professionals and non-professionals. The similarities and differences between these two levels of language offer many topics for research. The way science presents pieces of information has interested many writers (eg Halliday 1994, Skelton 1997). The differences between academic and popular presentations of science are discussed eg in Myers 1994. What is common to these studies is that they usually remain at the qualitative level. A computer corpus enables the researcher to combine qualitative and quantitative methods.

The idea of collecting a new computer corpus arose from the fact that available corpora were not ideal for studying how modal expressions (ie words and phrases expressing eg possibility and necessity) are used in different types of medical writing. Even though there are large corpora covering a wide variety of present-day English registers and text categories (eg the British National Corpus), they are not always suitable when studying a certain specific use of language. For some research interests, a specialized corpus is more useful than a large but more heterogeneous corpus.

Medicor is not the only corpus that aims to show the usefulness of scientific or medical corpora. Two other corpora deserve to be mentioned here. First, Juhani Norri (University of Tampere) is compiling a corpus of scientific texts published in the United States in the 1990s.[2] The corpus, which will comprise 1.2 million words, includes medical texts, but it covers other fields of science as well. It represents four levels of language: scientific journals, textbooks, magazine articles written by experts, and magazine articles written by journalists. *Medicor*, which covers only medical writing, is similarly organized but includes three additional text types: professional editorials, professional manual texts, and popular guidebook samples. In the main division of *Medicor*, magazine articles written by experts and journalists are grouped together. It is possible to treat them as different sections when using the corpus.

Second, the diachronic aspect of medical language is represented in a corpus compiled at the University of Helsinki by Irma Taavitsainen and Päivi Pahta.[3] This historical corpus and *Medicor* complement each other. Our aim is to combine them and to cover the time gap that now exists between them. The resulting corpus will represent medical writing from 1375 to 1997, from the late Middle Ages to the present day.

3 Medicor: main features

Medicor is characterized by three main features. First, it represents contemporary American English. The majority of the texts were published in the 1990s, and a minority dates from the 1980s. The time span of the book samples is from 1983 to 1996, whereas all journal, newspaper, and magazine articles were published in 1997. Second, the texts included in the corpus are complete articles, chapters, or manual entries. Writers' names, footnotes, endnotes, bibliographies, tables, etc. were omitted when the texts were edited into the corpus.

Third, *Medicor* is divided into sections representing different text types. The main division is into professional and popular texts. This distinction is based on the status of the intended readership. Professional texts are defined as texts addressed to professional readers, ie researchers, practitioners, and students of medicine. Popular texts, on the other hand, are defined as texts intended for the general readership, ie for those without medical training. The intended audience – and not the authors' background – is used as the criterion for classification; the writers of popular texts include both laymen and medical professionals. (The writers' background is presented in section 6 below.)

Professional and popular texts are further divided into the following text types (described in more detail in sections 4 and 5):

1. Professional texts: 1.1. textbooks, 1.2. handbooks, 1.3. research articles, 1.4. editorial articles.
2. Popular texts: 2.1. guidebooks, 2.2. newspaper/magazine articles.

Figure 1 shows the proportions of these text types. Table 1 gives the following information on the different text types of the corpus as well as on the corpus as a whole: number of texts, word count, and average word count in a text.

Table 1: A corpus of contemporary medical texts: number of texts, word count, and average word count in a text in the different text types

Text	Number	Word	Average word
type	of texts	count	count in a text
Professional texts	96	308,142	3,210
Textbook samples	11	104,702	9,518
Handbook samples	16	41,192	2,575
Research articles	33	114,432	3,468
Editorial articles	36	47,816	1,328
Popular texts	83	89,169	1,074
Guidebook samples	17	10,066	592
Newspaper/magazine articles	66	79,103	1,199
Total	179	397,311	2,220

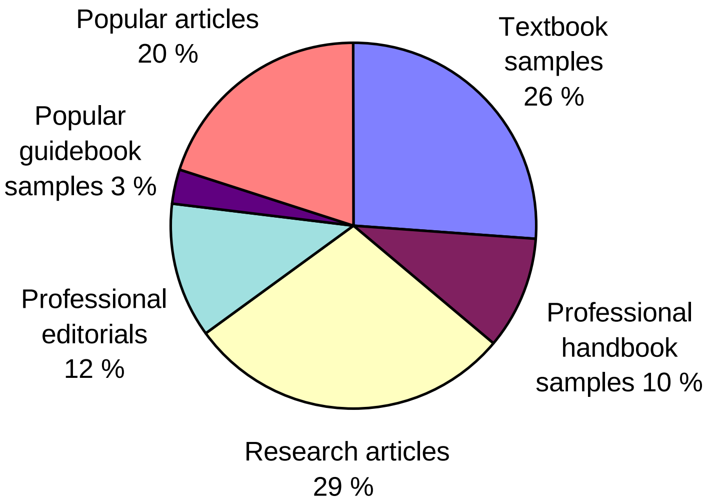

Figure 1: A corpus of contemporary American medical texts: proportions of different text types

4 Professional texts

4.1 Textbook samples

The textbook samples derive from books which are used in medical schools in various countries.[4] These books represent both scientific and clinical branches of medicine: while some of them present basic concepts and findings to a novice, others offer more practically oriented information to more advanced students. In addition to medical students, the books discussing the clinical aspects of medicine are used by practising doctors, and the works representing the scientific basis of medicine are of interest to researchers. Some of the books are used at the schools of dentistry and veterinary science as well.

4.2 Handbook samples

The handbook samples are chapters from a medical practitioner’s manual.[5] This manual has an international distribution, and is divided into chapters (written by different authors) each of which discusses a different disease or medical question. Its aim is to give concise information about different diseases, including their etiology, pathogenesis, symptoms, diagnosis, treatment, and prognosis. The samples included in the corpus represent different branches of medicine (eg neurology, cardiology, endocrinology, pulmonary diseases).

The manual is intended for both practitioners and advanced students of medicine. The reader is supposed to have background knowledge on the different branches of medicine, since the function of the text is to work as a reminder, rather than to introduce things to a novice. The manual is not a textbook of any branch of medicine, and it is not supposed to take the place of more thorough presentations. It supplements larger works and serves as a sourcebook when things have to be checked quickly.

4.3 Research articles

The research articles were published in established medical journals which have a worldwide distribution.[6] The journals represent different branches of medicine and cover both clinical and more theoretically oriented areas. The texts classified as ‘research articles’ range from rigorous biomedical basic research to applied research and scientifically less ambitious reports. All the research articles are original articles, presenting new findings based on the writers’ own study or research. Review articles were excluded from the corpus.

4.4 Editorials

Like the research articles, the editorials were published in academic medical journals. The editorials discuss various topics that the editors of the journals regard as important: new research results, methods of treatment, doctors’ training, the role of medical research, and ethical questions related to medical practice as well as to medical writing.

5 Popular texts

5.1 Guidebook samples

The guidebook samples were selected from a manual intended to help the general reader with medical questions.[7] The main part of the book discusses various diseases and disorders in an encyclopedic way from A to Z. Whereas some entries explain the functions of different organs as background information (these kinds of entries are not included in the corpus), others describe different diseases, discuss risk factors, and suggest forms of treatment. The samples included in the corpus present both common diseases (eg bronchitis) and less frequent disorders (eg endometriosis).

5.2 Popular articles

Some of the popular articles were selected from general newspapers and magazines, all having international distribution (*International Herald Tribune*, *Time*, *Newsweek*). Others derive from magazines intended for a more specified readership: one magazine specialized in science (*Scientific American*) and another specialized in health issues and practical aspects of medicine (*Prevention*). Both magazines are published monthly. Some of the popular articles give pieces of advice of the type ‘how to avoid flu,’ some present new biomedical research results to the general reader, and others discuss ethical questions related to medicine.

6 Writers

Information on the writers’ professional background is available in most of the publications. The majority of the writers in the professional category are doctors of medicine, some are PhDs, and those without a doctor’s degree are in a minority. The writers of the popular texts include both medical professionals (doctors, a nutritionist) and laymen (journalists). The writers represent both sexes.

In order to make the corpus linguistically more homogeneous, the writers had to fulfil two criteria for their texts to be included in the corpus. First, the writer whose name is mentioned first in the text must have an English name or at least an English first name. This criterion aims to minimize non-native writer influence. Unfortunately, this criterion could not be used with all popular texts, as the writer’s name is missing in some of them. However, it can be expected that eg professional newspaper reporters’ writing represents native-level language.

Second, a text was included in the corpus only if the writer (or at least the first writer) was affiliated to a US hospital, university, or institution. This was done in order to minimize the variation caused by possible regional differences (see section 7 below). Naturally, this criterion could be used only when background information on the writer was available in the publication.

7 Why American?

The corpus represents American books and journals, and studying medical language through American English texts is well grounded. First, English is the language that enables communication in the medical researchers' international and heteroglossic sphere. Second, American texts in particular have a large distribution: American books are used in medical schools in many countries, and American scientific publications convey new findings worldwide. In addition, many American journals that are intended for the general readership have a wide international distribution.

A different kind of reason for choosing only American texts was that the corpus was not designed to be a means of comparing possible regional differences within medical writing. As only American texts are included, the user of the corpus does not have to speculate whether observed differences between the texts of the corpus are caused by regional variation (eg British vs. American English) or text type variation. To cover both regional and text type variation scales the corpus should have been of a larger size. The corpus can be supplemented later by British English texts.

8 Conclusion

Medicor is a new corpus of contemporary medical texts. At present, it is not coded, but it will be provided with a coding system giving information on the source, text type, and writers (sex, medical professional/ non-professional). Grammatical and syntactic tagging will also be added to the corpus using an automatic grammatical parser.

Medical discourse offers interesting research opportunities. They include, for example, the way language is used to create hypotheses, the differences between professional and popular levels of language, and the relationship between linguistic form and scientific background knowledge. *Medicor* is designed to serve these kinds of research interests. When completed, it will provide material for researchers of medical language as well as for teachers and students of English.

Notes

[1] As I am writing my PhD on modality in medical texts, I needed a selection of texts representing different types of medical writing. For the time being, the corpus is only used for this work, but later it will be distributed for public use.

[2] For details, see Norri and Kytö 1996.

[3] For further information, see Taavitsainen and Pahta 1997.

[4] Examples of the sources:
Andreoli, Thomas et al (eds). 1993. *Cecil essentials of medicine*. 3rd ed. Philadelphia: Saunders.
Guyton, Arthur and John Hall. 1996. *Textbook of medical physiology*. 9th ed. Philadelphia: Saunders.
Jorde, Lynn et al. 1995. *Medical genetics*. St. Louis: Mosby.

[5] Robert Berkow et al (eds). 1987. *The Merck manual*. 15th ed. Rahway NJ: Merck Sharpe & Dohme.

[6] E. g. *American Journal of Pathology, Annals of Neurology, Archives of Ophthalmology, Journal of Pediatrics*.

[7] *Family medical guide: The illustrated medical and health advisor*. 1983. By the editors of Consumer Guide with medical consultants Ira Chasnoff, Jeffrey Ellis and Zachary Fainman. New York: Morrow.

References

Halliday, M. A. K. 1994. The construction of knowledge and value in the grammar of scientific discourse, with reference to Charles Darwin’s *The Origin of Species*. In M. Coulthard (ed) *Advances in written text analysis*, 136–156. London: Routledge.

Myers, Greg. 1994. Narratives of science and nature in popularizing molecular genetics. In M. Coulthard (ed) *Advances in written text analysis*, 179–190. London: Routledge.

Norri, Juhani and Merja Kytö. 1996. A corpus of English for specific purposes: Work in progress at the University of Tampere. In C. Percy et al (eds) *Synchronic corpus linguistics: Papers from the sixteenth international conference on English language research on computerized corpora (ICAME 16)*, 159–169. Amsterdam: Rodopi.

Skelton, John. 1997. The representation of truth in academic medical writing. *Applied Linguistics* 18:2:121–140.

Taavitsainen, Irma and Päivi Pahta. 1997. Corpus of Early English medical writing 1375–1750. *ICAME Journal* 21:71–81.

A Frequency Dictionary of Contemporary American English: word sketches, collocates, and thematic lists is an invaluable tool for all learners of American English, providing a list of the 5,000 most frequently used words in the language. The dictionary is based on data from a 385-million-word corpus evenly balanced between spoken English (unscripted conversation from radio and TV shows), fiction (books, short stories, movie scripts), more than 100 popular magazines, ten newspapers, and 100 academic journals for a total of nearly 150,000 texts. All entries in the rank frequency list feature the Medicor: A corpus of contemporary American medical texts. May 1998. Minna Vihla. This paper introduces a new computer corpus which will enable quantitative study of medical discourse. The new corpus, Medicor, contains contemporary American medical texts, and its size is 397,311 words. It is being compiled at the University of Helsinki by Minna Vihla. Read more.