

THE (VERY) LONG HISTORY OF CORPORA, CONCORDANCES, COLLOCATIONS AND ALL THAT

Michael Stubbs

In the development of academic disciplines, important ideas are often proposed, forgotten, and then rediscovered much later, when they are connected to other ideas in a way which reveals their significance. I give examples of ideas which are often thought of as quite modern, although they have a very long history:

- using corpora in constructing dictionaries and language teaching materials
- using concordances as data for textual exegesis and information retrieval
- using collocations as evidence of word meaning.

In all three cases the theoretical significance of the ideas became clear only after improved techniques of visualization allowed patterns to be seen in complex non-numerical data. The many accounts of the history of corpus study are all partial in different ways, and my account here is also a mere sketch of the origins of some major ideas. My main focus is on the history of attempts to visualize patterns in texts and how these attempts led, over a long period of time, to important linguistic concepts.

First, important ideas are often proposed, forgotten about, and then rediscovered, independently, sometimes hundreds of years later. In the natural sciences, so it is said, "science destroys its past" (Kuhn 1969). But a survey of computing in the humanities makes the point that:

It is good every so often to remind ourselves that we are part of a line of historical development. It is particularly appropriate to do this in times which are described as new, ground-breaking, unique ... [W]e may [...] labour under the misconception that what we are doing is so radically different from before. (Fraser 1996.)

Second, the reason why many precursors of important ideas sank without trace is that they were not connected to other ideas, and so their full implications could not be seen. This process of connection sometimes took hundreds of years.

Third, corpus linguistics, in its modern sense, could obviously develop only with computational help. However, it is not technology alone which leads to theory,

but the techniques of visualization which the technology makes possible, and the importance of visualization was recognized long before computers were available.

Fourth, practice came before theory. Real world problems required practical applications which, in turn, led to theoretical developments (often helped by visualization techniques). Many important concepts originated in textual exegesis, language teaching, lexicography, and information processing.

Fraser, Michael. 1996. A hypertextual history of humanities computing: the pioneers. <<http://info.ox.ac.uk/ctitext/history/pioneer.html>>

Kuhn, Thomas S. 1969. Comment on the relations of science and art.

Comparative Studies in Society and History. 11: 403-12. 14 April 2015.
