Towards a methodology for corpus-based studies of linguistic change

Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish

Jørg Asmussen
Society for Danish Language and Literature
Det Danske Sprog- og Litteraturselskab
DSL
The Korpus 2000 Project, www.dsl.dk/korpus2000
Christians Brygge 1, DK-1219 Copenhagen K
ja@dsl.dk

Abstract

Corpora serve as a widely accepted base for synchronic descriptions of language. Yet easily accessible general corpora that enable diachronic descriptions of language are still quite rare, the Korpus 90 and Korpus 2000 Corpora of Danish being one exception.

Korpus 90 and Korpus 2000 were both designed and compiled at the *Society for Danish Language and Literature (DSL)*. Korpus 90 comprises text material from the period 1983-1992 and was compiled in the early 1990s. Korpus 2000 is a recently compiled corpus holding text material from the period 1998-2002. The joint web-based query interface of the two corpora enables immediate comparative studies.

This paper first gives a very brief introduction to the background of the two corpora before focusing on examples of contrastive observations and their possible diachronic interpretations - and misinterpretations. The examples cover frequencies (new words, vanishing words), the inflectional and collocational behaviour of certain words, and their connotations. An example illustrating syntactical differences is also briefly sketched. The paper then discusses whether these observable differences reflect real changes in the Danish language, or whether they reflect the probable fact of differently compiled - and thus perhaps incomparable - corpora.

Finally, the paper proposes some prerequisites for a methodology of comparative corpus investigation and the determination of diachronic corpus similarity. In this context, the concept of invariant textual features will be introduced.