Henning Bergenholtz

- Centre for Lexicography
  Faculty of Social Sciences
  University of Aarhus, Denmark
- Department of Information Science
  Faculty of Engineering, Built Environment & IT
  University of Pretoria, South Africa
- Department of Afrikaans and Dutch
  Faculty of Arts and Social Sciences
  University of Stellenbosch, South Africa

Title

- A corpus analysis is a superfluous ceremony and a complete waste of your time and the government's money

- **Not my opinion!**

Three attitudes

- A corpus analysis is a waste of time
- An electronic corpus is necessary for linguistic and also for lexicographic projects; without the use of a corpus the research is not scientific
- Many linguistic and lexicographic projects must have a corpus as empirical basis, but not all

A corpus analysis is a waste of time

That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text. (Robert Lees in a discussion 1962, quoted in Francis 1979:110)

A corpus analysis is a waste of time

Francis, W. Nelson: Problems of Assembling and Computerizing Large Corpora. In: Henning Bergenholtz/ Burkhard Schaeder (Hrsg.): *Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora.* Königstein/Ts.: Scriptor, 1979, 110–123.

A corpus analysis is a superfluous ceremony

Itkonen, Isa 1976: Was für eine Wissenschaft ist die Linguistik eigent-lich? In: Dieter Wunderlich (Hrsg.): *Wissenschaftstheorie der Lin-guistik.* Kronberg: Athenäum, 56–76.

Page 65: Im Zusammenhang mit grammatischen Beschreibungen ist die Heranziehung eines Korpus von tatsächlichen Äußerungen also eine überflüssige Zeremonie.

Linguists use their knowledge of the language to create a corpus

Linguists use an existing corpus or a corpus created from introspection to discover criteria for categorization or the formulation of rules.

Greenbaum, Sidney: Corpus Analysis and Elicitation Tests. In: *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research,* ed. by Jan Aarts and Willem Meijs. Amsterdam: Rodopi, 1984, 193–201.

Self-made examples and collocations

HB calls self-made examples ”Linguistic Poetry” comparing the examples in a German dictionary where 33% of the self-invented examples look like:

- **beginnen** Die Hausfrau beginnt damit, das Geschirr abzutrocknen.
- **begegnen** Wir begegnen dem Vater mit Achtung.
- **beobachten** Das Kind beobachtet es genau, wie sein Vater das Auto repariert.
- **besorgen** Die Hausfrau hat Milch besorgt.

HB is not against the use of corpus, but …

- The first ”big” German corpus: **Limas-Korpus**. Bonn 1975.
- One of the first big Danish corpora: **DK87-90**. Århus 1991.
- Several specialised language corpora, e.g. **Gene Technology Corpus**. Århus 1991.
- A lot of corpus contributions, e.g. Henning Bergenholtz/ Joachim Mugdan: Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität; in: *Computional Linguistics. Computerlinguistik. An International Handbook on Computer Oriented Lan-guage Research …*, hrsg. von Istvàn S. Bàtori, Winfried Lenders, Wolfgang Putschke. Berlin/New York: de Gruyter 1989, 141-149.

HB is not against the use of corpus, but …

HB was one of the editors of – I think – the first book about corpus in Europe:

Henning Bergenholtz/Burkhard Schaeder (Hrsg.): *Empirische Textwissen-schaft. Aufbau und Auswertung von Text-Corpora.* Königstein/Ts.: Scriptor 1979.

This book is not well-known today because of its German title, but most of the contributions were in English, e.g. the contributions by W. Nelson Francis, Stig Johansson, Randolph Quirk, Jan Svartvik, Roger G. van de Velde, Bjarne Ulvestad

Many linguistic and lexicographic projects must have a corpus as an empirical basis, but not all

My thesis is therefore somewhere in the middle of **we do not need corpora** and **habeas corpus**, the almost religious belief that a corpus is needed for every kind of linguistic and lexicographic work, without a corpus you will not get into heaven.

Examples, collocations

If you do not want to make the error of linguist poetry, you can ONLY find examples and collocations taken out of real texts; this will in most cases mean that you need an electronic readable corpus.

Personally I do not use the different programs for an automatic search. The results are of different reasons not impressive. But that is not the point here.

In all the dictionary projects I have been part of in the last 25 years, we have only had examples and collocations taken out of a corpus.

Meaning items

In general language lexicography you can for the moste lemmas only formulate the meaning items and here too make the polysemy decisions, if you use the examples in a corpus.

See e.g. Bergenholtz/Agerbo 2014 (in Lexicographica *30*, 488-510)

Frequency

Of course you can only find data about frequencies by an analysis of one or more corpora. But such results are often used in dictionaries in a not relevant way, especially in English dictionaries with marks for frequent, very frequent, not frequent etc. But why?

For user needs during text reception? Text production?

For a cognitive function: Knowledge. OK, but

Need for a corpus, but not an e-corpus

No examples from lexicography but from grammar research.

The German pronoun es:

Er liebt es, dass … Er liebt, dass …

Er sagt es, dass … Er sagt, dass …

In total we used a reading corpus with 40 mio. running words for our research.

No need for a corpus

- Lemma selection for the Danish Music Dictionary:
- We did not use any kind of corpus for the work with the database for music dictionaries, neither for the lemma selection. We used instead the lemma stock in already existing music dictionaries + indexes in music handbooks
- In a discussion at the Asialex symposium Mr. Kilgariff attacked this solution telling that terminology experts told him that a corpus analysis is needed. I am professor for bilingual specialized lexicography. He could have asked me.

No need for a corpus

In some kind of specialized dictionaries we cannot use a corpus to get the meaning items, and a specialized corpus neither. We need specialized knowledge. So we have done it in the music dictionaries, the molecular biology dictionaries and in the accounting dictionaries. Would you use a corpus to make the meaning entry in a linguistic dictionary, e.g. for the term **adjective**?

Perhaps you would make a Google-search, but I don't think you should accept the results uncritically. But so argued the same Kilgariff criticizing our accounting dictionaries.

Google-search or corpus analysis for a meaning entry in an accounting dictionary?

- What is *deemed cost* really? The question came up by reading Kilgariff (2012: 27), who criticizes THE ENGLISH-SPANISH ACCOUNTING DICTIONARY for making a too long definition:
- "Deemed cost is an amount used instead of cost or depreciated cost at a specific date. Any following amortisation or depreciation is made on the assumption that the enterprise initially recognised the asset or liability at a cost equal to the deemed cost."

What is *deemed cost* really?

Kilgariff (2012) has instead a much easier solution, one every lexicographer should use in his running work, he writes:

"Because they could have found a shorter and better one by a Google search: Surrogate for cost at a given day."

Bergenholtz et al.: What is *deemed cost* really?

No, if you are an expert, you know the international norm, from IFRS, the official international standard commission. Here you see that *Deemed costs* have to be recognised and noted in the company's books as such before the term can be used, e.g. by *amortisation*.

A linguist cannot know that, but he cannot decide neither if a use in his corpus or by a Google-search is correct or not.

One example more

We did not use a corpus for the definition of **gene** (in a dictionary for semiexperts):

A gene is a DNA sequence encoding a protein. tRNA or rRNA. For eukaryotes a gene can also be defined as a transcribed DNA sequence or transcriptum unit. In prokaryotes two or more proteins are often encoded in the same transcription unit, and such a transcription unit plus its associated regulatory sequences is termed an operon.

One example more

In another dictionary for laymen we have the following meaning items, only the second one is made outgoing from a corpus analysis:

1. hereditary systems related to the chromosomes in female or male sex cells in human beings and animals
2. tendency or disposition for to like to do something

Corpus linguistics

- I never use the term corpus linguistics. I do not use the term intuition linguistics or survey linguistics. Using a corpus is using a certain empirical basis. Not more, it cannot and will not be a linguist discipline like sociolinguistics or grammar.
- Corpus as an empirical basis is needed in many kinds of linguistic research processes, but not always.

Corpus lexicography

I never use the term corpus lexicography. I do not use the term intuition lexicography or survey lexicography. Using a corpus is using a certain empirical basis. Not more, it cannot and will not be a lexicographic discipline like specialized lexicography or general language lexicography.

Corpus as an empirical basis is needed in many kinds of lexicography processes, but not always.

Corpus lexicography

"The Aarhus School doubts the relevance of corpora for lexicography (explicitly, in the concluding chapter, p309). But you need corpora to get the facts right."

Kilgariff (2012, 29)