The Use of Text Corpora in Lexical Research

Stefan Engelberg

Workshop, Universitatea din București, November 2008

http://www.ids-mannheim.de/ll/lehre/engelberg/Webseite_CorpLex/CorpLex.html

engelberg@ids-mannheim.de

Content

1. Empirical linguistics
2. Text corpora and corpus linguistics
3. Concordances
4. Application I: The German progressive
5. Part-of-speech tagging
6. Frequency analysis
7. Application II: Compounds
8. Co-occurrence analysis
9. Application III: Word senses in lexicography
10. Keyword analysis

Corpus analysis software
AntConc
COSMAS II
Corpus Browser (Leipzig)
DWDS corpora & analysis
KWIC Finder
Empiricism in philosophy

**Empiricism**

Empiricism is an epistemological position that emphasizes that nothing can be known unless it is substantiated by information we gain from the senses.

- All concepts are based on experience.
- All true statements are based on experience or are logically derived from statements that are based on experience.
- (E. g., TIME is a concept derived from the nature of observable events.)

**Rationalism**

Rationalism is an epistemological position that emphasizes that knowledge can be gained from reasoning based (at least partly) on concepts that are given a priori.

- Basic concepts are innate.
- (Some) true statements can be made without recurrence to experience.
- (E. g., TIME is an innate concept, prior to experience.)
1 Empirical linguistics

1.1 Empiricism

Empirical research

Empirical research bases all its findings on direct or indirect observation.

• It proceeds inductively from data to theory formation.
• It tends to be data-driven, i.e., based on data that are analyzed with few preconceived assumptions about their structure.
• (E.g., in linguistics, it is rather based on extensive corpus linguistic research or psycholinguistic experiments, than on single, subjective grammaticality judgements.)

Theory

A theory is a set of connected statements about a part of reality. The set of statements includes descriptions, explanations, and laws. A theory is falsifiable and it allows to make predictions about phenomena which have not been observed before.

Stefan Engelberg (IDS Mannheim), Workshop „Corpora in Lexical Research“, Bucharest, Nov. 2008 [Folie 5]
1 Empirical linguistics
1.2 The scientific process

The BEGINNING of a scientific investigation

Wondering …

„Wonder […] and not any expectation of advantage from its discoveries, is the first principle which prompts mankind to the study of Philosophy, of that science which pretends to lay open the concealed connections that unite the various appearance of nature; and they pursue this study for its own sake, as an original pleasure or good in itself, without regarding its tendency to procure them the means of many other pleasures. (Section III).“ (Adam Smith, *The History of Astronomy*, Section III)

The WAY of empirical investigations

„Wondering“ wondering about a particular state of affairs; a „Why“-question

Inquiry
examination of current theories about the phenomenon

Hypothesis
explicit formulation of a statement about the phenomenon, that can be justified and is presumed to be true

Research design
Design of an empirical investigation

Data acquisition
questionnaire, psycholinguistic experiment, corpus study, …

Data processing
transcription, collection of data in data bases, clearing up of corpus data, …

Data analysis
classification, typology, qualitative analysis, …

Theory
interpretation of data, formulation of theoretical statements

falsification
acquisition of further data

Stefan Engelberg (IDS Mannheim), Workshop „Corpora in Lexical Research“, Bucharest, Nov. 2008 [Folie 7]
Two types of empirical investigations

**Testing hypotheses** (cf. previous slide):
Starting point: a hypothesis derived from a theory, a preparative study, or a personal conviction.
Result: the confirmation or falsification of the hypothesis.

**Exploring hypotheses:**
Starting point: the investigation and systematization of data, from which hypotheses about the general nature of the phenomena can be derived.
Result: scientific hypotheses.

1 Empirical linguistics
1.2 The scientific process

**Data** are state of affairs that can be observed and that are used in order to argue for or against theoretical assumptions.

Linguists use different types of data which are gained in very different ways:

**(I) Introspective grammaticality judgements**

As a native speaker of German I judge the following sentences as indicated:

* Peter hilft seinen Freund
  
  Peter hilft seinem Freund
1 Empirical linguistics
1.3 Types of data in linguistics

(II) Grammaticality judgements on the basis of operational tests

Verb phrases of a particular semantic type (so-called accomplishments) can be
determined by a span-of-time test; they can be combined with an adverbial of the type
in x Tagen / Stunden / Minuten:

(1) Er hat den Wagen in drei Stunden repariert.
(2) *Er hat den Mechaniker in drei Stunden gehasst.


1 Empirical linguistics
1.3 Types of data in linguistics

(III) Graded grammaticality judgements on the basis of larger number of native speakers

On an acceptability scale from 1 (entirely ok) till 5 (unintelligible) 180 subjects judged
the following sentence as 2,9 on average:

Das echte Überraschen der Kinder beim Anblick der Eltern war niemandem entgangen.

Blume, Kerstin (2004): Nominalisierte Infinitive: Eine empirisch basierte Studie zum Deutschen. Tübingen:
Niemeyer. S. 69.
1 Empirical linguistics
1.3 Types of data in linguistics

(IV) Results of psycholinguistic experiments

If two words are presented in succession and subjects have to decide as fast as possible whether the second word is a word of German, they solve this task faster in case (2) than in case (1):

(1) Fahrrad – Doktor.
(2) Krankenschwester – Doktor.


(V) Scans on the basis of imaging techniques

When trilingual subjects were asked to correct grammatical errors they exhibit brain activity in different areas of the brain (language 1: orange; language 2: blue; language 3: green). The left image shows the activities in the Broca area of a subject who learnt two of the three languages befor the age of three (one network for the two languages), the right image the activities of a subject who learnt two of the three languages after the age of ten (one network for each language):

(VI) Results of electroencephalography (EEGs)

The EEG measurement relative to a particular event (event-correlated potential), namely the presentation of the sentences (1) through (3), shows differences in the area of the N400 component:

1. The pizza was too hot to eat.
2. The pizza was too hot to drink.
3. The pizza was too hot to cry.


Dialogue between a patient (40 years old, housewife, stroke) and a (female) examiner:

P: äh / is vor drei Jahren Unfall / äh / im Auto passiert /
U: ein Unfall? /
P: ja / im / äh / Stuttgart /
U: ahja /
U: Sind Sie ohnmächtig dann geworden?
P: Nein! / mein Mann / äh / is / äh / ich in Bett / ge / gegegangen und / äh / und / äh / der Arzt gerufen / und / äh / aufschreiben / gar nix! / weg! / und dann dann / äh / ohnma / ne nich / ohnmächtig! / das is nicht / äh / so / äh / so gef / äh / ganz / ganz / weit weg / Gedanken // ja und dann ... /
(VIII) Utterances of (healthy) speakers from electronic text corpora

Concordance for *besteht*, created by AntConc, version 3.2.1w, on the basis of part of the German newspaper corpus within the Leipzig Corpus Collection.

(IX) Statistical measurements

Statistical computation of the collocation potential of the word *bestehen* on the basis of the „Deutsches Referenzkorpus“; excerpt from the CCDB (co-occurrence database).