Statistics for linguists

Holger Diessel
holger.diessel@uni-jena.de

Qualitative vs. quantitative research

Research in linguistics can be based on qualitative and quantitative data.

For a long time linguistic research was almost exclusively qualitative.

Qualitative research

(1) a. *Himself saw me.
b. John$_i$ saw himself$_i$.
c. *John$_i$ saw himself$_j$.
d. *I$_i$ want John to paint myself$_i$.

(2) a. He saw me.
b. *John$_i$ saw him$_i$.
c. John$_i$ saw him$_j$.
d. I$_i$ want John to paint me$_i$.

(3) a. John saw me.
b. *The man$_i$ saw John$_i$.
c. The man$_i$ saw John$_j$.
d. *I$_i$ wants John to paint Mary$_i$.

Binding principles

1. An anaphor is bound in its governing category
2. A pronoun is free in its governing category
3. A referential expression is free

The empirical turn

The empirical turn in linguistics is accompanied by theoretical changes:

In structuralist and generativist theories of language, language is seen as a closed deductive system (similar to mathematics or logic). In such a view of language, a single sample sentence is sufficient to draw a valid conclusion.

In usage-based theories of language, language is seen as a dynamic system consisting of fluid categories and gradient constraints. In such a view of language, researchers have to work with quantitative data and probabilities.

Example of quantitative research

A child language researcher wants to find out if girls are more advanced in word learning than boys. She collects the following data from 20 boys and 20 girls at the one-word stage (at which parents are usually aware of all words their children know):

Average number of words girls know at the age of	1;6:	56.3
Average number of words boys know at the age of	1;6:	53.9

Do these data support the researchers hypothesis?

In order to answer this question, we have to consider these issues:

- Is the sample representative?
- Is the sample large enough?
- Is the difference between 56.3 and 53.9 large enough?

Syllabus

What you will learn:

1. Basic concepts of statistical analysis
2. Introduction to SPSS

The readings: Agresti and Finlay 2009

Data in linguistics research

Linguistic data:

- Sample of written texts
- Sample of transcripts of spoken language
- Large electronic corpora
- Experiments (psycholinguistics)
- Questionnaire (socio-linguistics)
- Dictionaries (e.g. OED)
- Diary data (child language)
- Reference grammars (linguistic typology)
- Videos (pragmatics)

Is introspective data useful?

Data in linguistics research

Empirical research in the social sciences has three general goals:

- Description
- Prediction
- Explanation

Two types of empirical studies:

- Explorative
- Hypothesis testing

Sampling

Empirical (quantitative) research never really proves the truth of anything; it provides evidence against or in favor of a particular hypothesis. Empirical methods help us to evaluate this evidence. Evaluating an empirical study involves (at least) the following issues:

- Sampling procedure
- Sample size
- Coding
- Experimental task and experimental design
- Statistical analysis

Sampling

Empirical research is almost always based on samples (i.e. subsets of the true population). This is a risky procedure.

Example: It is commonly assumed that passive sentences are more difficult or more complex than active sentences. Does the difficulty/complexity correlate with the length of active and passive sentences? -> How can we investigate this question?

- Corpus data
- Experiment

The researcher finds out that passive sentences are indeed somewhat longer than active sentences in both a large electronic corpus and a production experiment. Does this prove that his hypothesis is true?

General criteria for empirical research

There are three general criteria for good empirical research:

- Generalizability: Can we generalize from the sample to the true population?
- Reliability: Is the way the data has been measured and analyzed precise and correct? -> Is the study repeatable?
- Validity: Does your study really measure what it is supposed to measure. Is the experimental task etc. appropriate?

Occam's razor: Your account should be minimal: Don't explain your findings with more generalizations than necessary.