Friedrich-Alexander-Universität
Philosophische Fakultät und
Fachbereich Theologie

How Random is a Corpus?

Topic 1.2.A Methodological issues and statistical analysis

Xinyao Lu **Nathan Dykes**[1]

[1]Department of German Language and Literature, Chair of Computational Corpus Linguistics

12.07.24

1. Topic Orientation

1.1 Seminar Review

1.2 Procedures of a Corpus Linguistics research

1.3 Topic key word

2. Randomness and Non-Randomness

2.1 Source of randomness

2.2 Sources of non-randomness

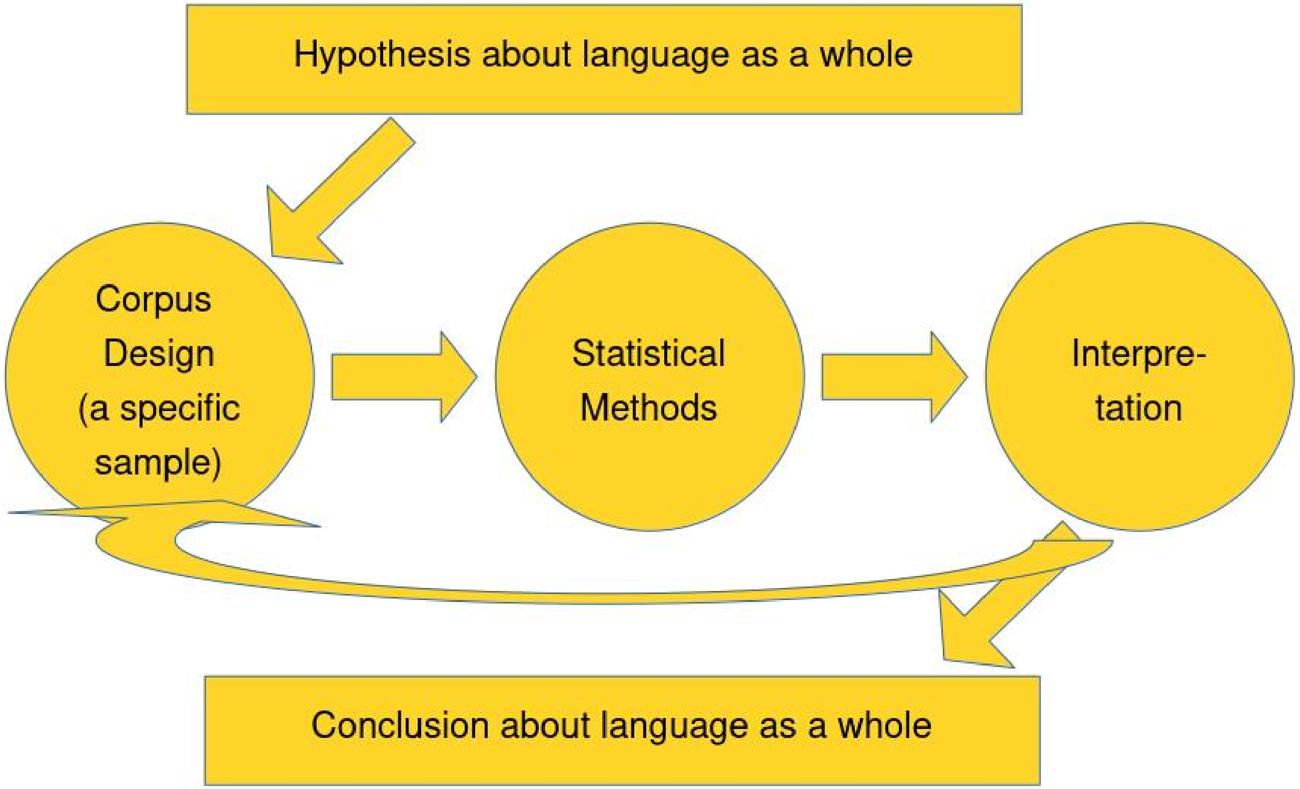
3. Conclusion

Seminar Review

What did we learn in this seminar?

- **Significance testing (Statistical methods)**
  - Null hypothesis, p-value, chi-square test
- **Corpus design**
  - Social media corpora
  - Spoken language corpora
- **Application of Corpus Linguistics (Results Interpretation)**
  - Critical discourse analysis
  - Translation studies
  - Pragmatic studies

Procedures of a Corpus Linguistics research



Hypothesis about language as a whole

Corpus Design (a specific sample)

Statistical Methods

Interpre-tation

Conclusion about language as a whole

Topic key word

Evert: Reflection of Statistical Methods in Corpus Linguistics

- *why should we apply statistical methods (based on the **random sample** model) at all?* (in corpus linguistics)
  - Evert, Stefan. "How Random is a Corpus? The Library Metaphor" Zeitschrift für Anglistik und Amerikanistik, vol. 54, no. 2, 2006, pp. 177-190.
- What does **Randomness** actually mean in Corpus Linguistics?
- Sources of **(non-)Randomness**: Corpus design
- Validating the **(non-)Randomness**: Results interpretation

1. Topic Orientation

1.1 Seminar Review

1.2 Procedures of a Corpus Linguistics research

1.3 Topic key word

2. Randomness and Non-Randomness

2.1 Source of randomness

2.2 Sources of non-randomness

3. Conclusion

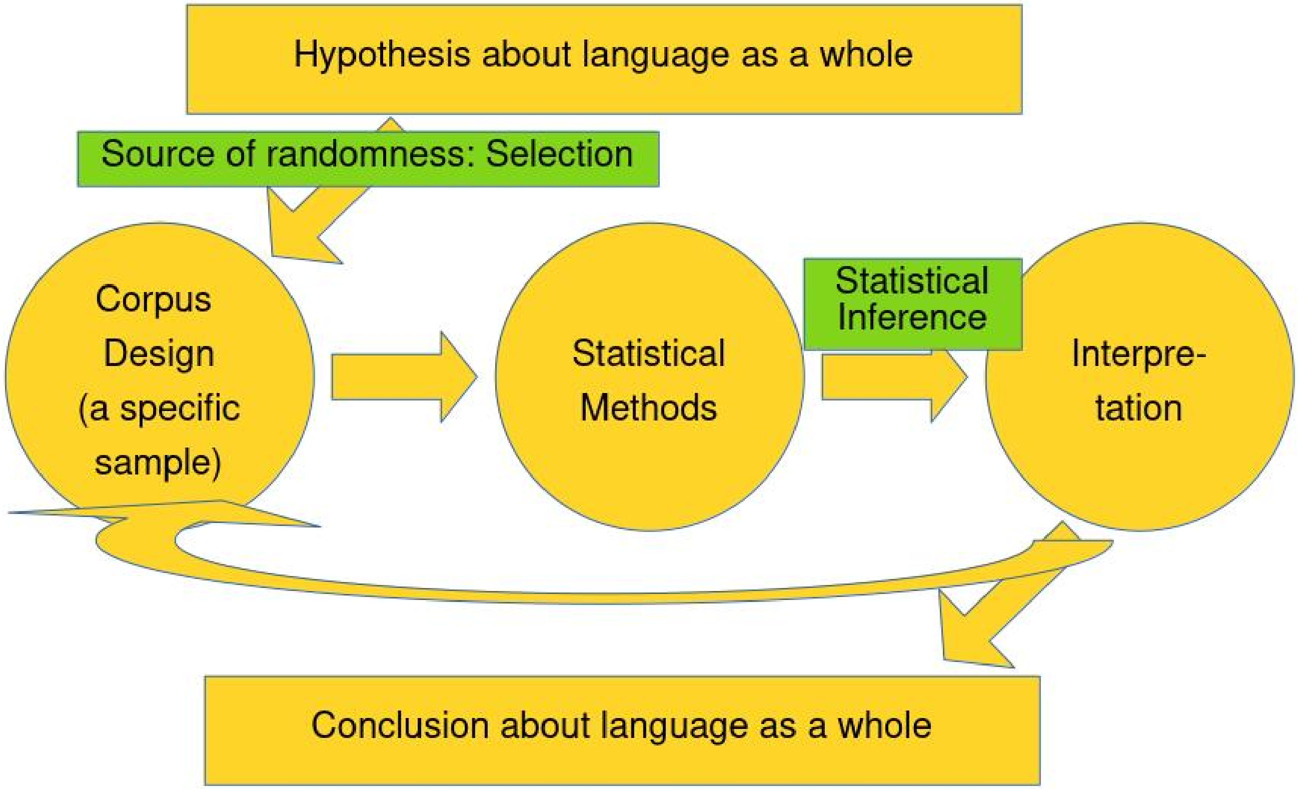
Source of randomness

The library metapher

- **A seeming conflict**
  - *statistical methods ... operate on random samples*
  - *very little (in language) is left to chance*
- **The library metapher**
  - *... there is nothing random about the text in the library: every senetence was produced for some specific purpose.*
  - *The selection of a particular corpus - picking an arbitrary book from one of the shelves*
  - *It is this choice which introduces an element of randomness into corpus frequency data.*
- **Why randomness?**
  - Statistical inference: Representativeness to (sub)language as a whole

Source of randomness

Illustrated in the procedures



Hypothesis about language as a whole

Source of randomness: Selection

Corpus Design (a specific sample)

Statistical Methods

Statistical Inference

Interpretation

Conclusion about language as a whole

Sources of non-randomness

Still not (completely) random?

- ***Balanced* samples**
  - **External** source of non-randomness
  - An issue of subjective language interpretation

- **The unit of sampling** vs. **the unit of measurement**
  - **Internal** source of non-randomness
  - Clustering effect: *a tendency to lump together*
  - Language is NOT a bag of random words

An evidence of internal non-randomness

Underestimated variation

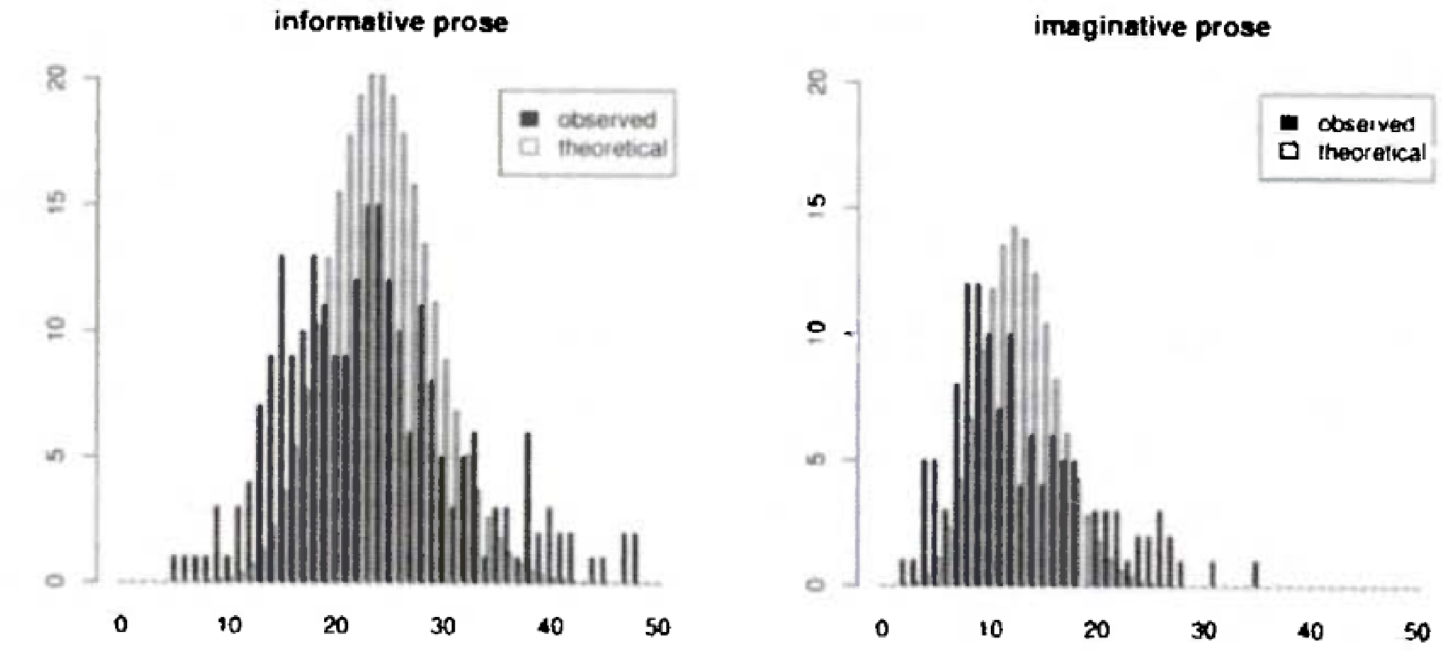


Figure 1: Empirical distribution (dark bars) of the frequencies of passive verb forms in the documents of the Brown corpus vs. theoretical binomial distribution (light bars) predicted by the random sample model

1. Topic Orientation

1.1 Seminar Review

1.2 Procedures of a Corpus Linguistics research

1.3 Topic key word

2. Randomness and Non-Randomness

2.1 Source of randomness

2.2 Sources of non-randomness

3. Conclusion

Conclusion

Things that statistics won’t tell you.

- **Statistical methods and corpus design**
  - Statistics don’t show the quality of corpus design.
  - Conversely, the selection of texts provides the randomness (i. e. representativeness), which is the prerequisite of all statistical methods.
- **Statistical methods and interpretation**
  - Two sources of randomness may reduce the validity of statistical reference: Imbalanced sampling and clustering effect.
  - Real linguistic data usually have larger variation (an index of non-randomness) than theoretical expectation.

Thanks for your attention

How does the elephant look like?

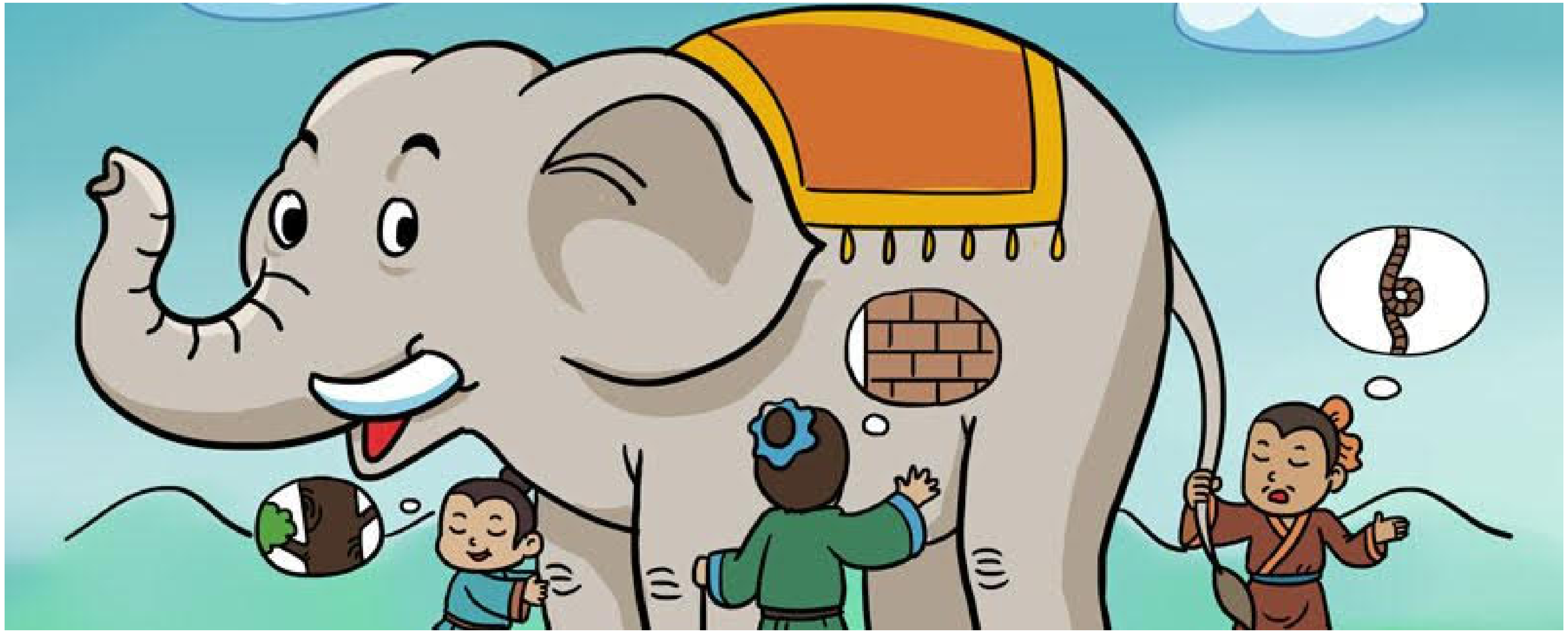


Figure: Blind people touch an elephant

References

[BE09] M. Baroni and S. Evert. “36. Statistical methods for corpus exploitation”. In: *Volume 2*. Ed. by A. Lüdeling and M. Kytö. Berlin, New York: De Gruyter Mouton, 2009, pp. 777–803. DOI: doi:10.1515/9783110213881.2.777.

[Eve06] S. Evert. “How Random is a Corpus? The Library Metaphor”. In: *Zeitschrift für Anglistik und Amerikanistik* 54.2 (2006), pp. 177–190. DOI: doi:10.1515/zaa-2006-0208.

[Ste20] A. Stefanowitsch. *Corpus linguistics: A guide to the methodology*. Textbooks in Language Sciences. Language Science Press, 2020.