CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

Introduction to Text Corpora and Their Applications

Corpora in contrastive linguistics and translation studies

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

OUTLINE:

1. LECTURE

- corpus-based contrastive linguistics
- corpus-based translation studies (CTS)
  - the so-called translation universals (TU)
  - types of corpora in CTS

2. SEMINAR

- reading (Andrew Chestermann): *Hypotheses about TU*
- S-universals and T-universals: how can they be studied?

LECTURE

Corpus-based contrastive linguistics

The beginnings of a new era

the 90s: comparison of languages on the basis of language corpora and use of corpus linguistics methods

- combining the methodological advantages of CL and the possibility of contrasting parallel texts in two and later even several languages
- greater accuracy and detail in research al all levels of description (from grammar and lexis to discourse)
- implications for other areas: language teaching, lexicography, translation studies and CAT

ENPC by Stig Johansson

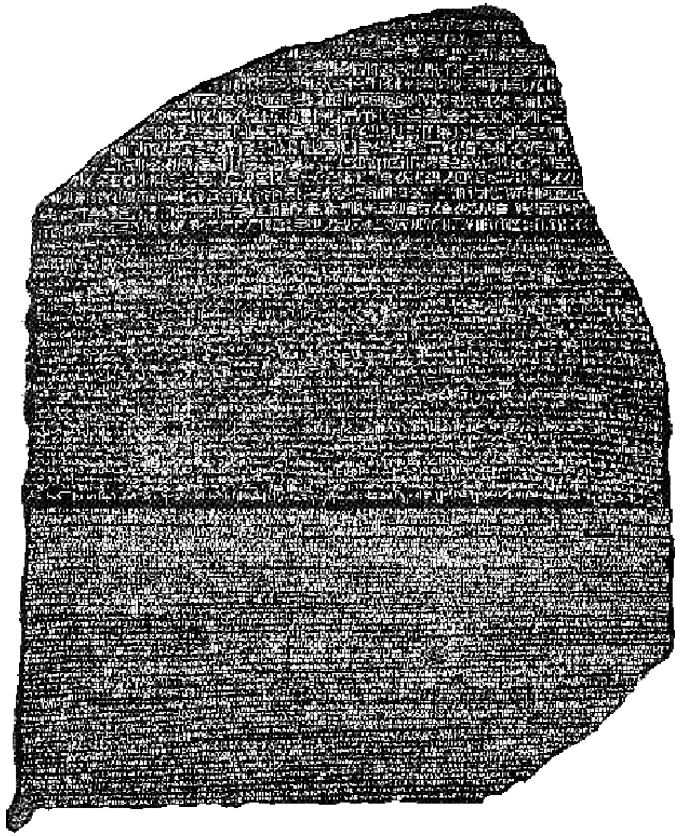
- Johansson and his team: English-Norwegian Parallel Corpus
  - unique project of the time
  - bidirectional translation corpus consisting of comparable English and Norwegian original texts and their translations into the other language

- „parallel corpus“
  - according to the Rosetta stone

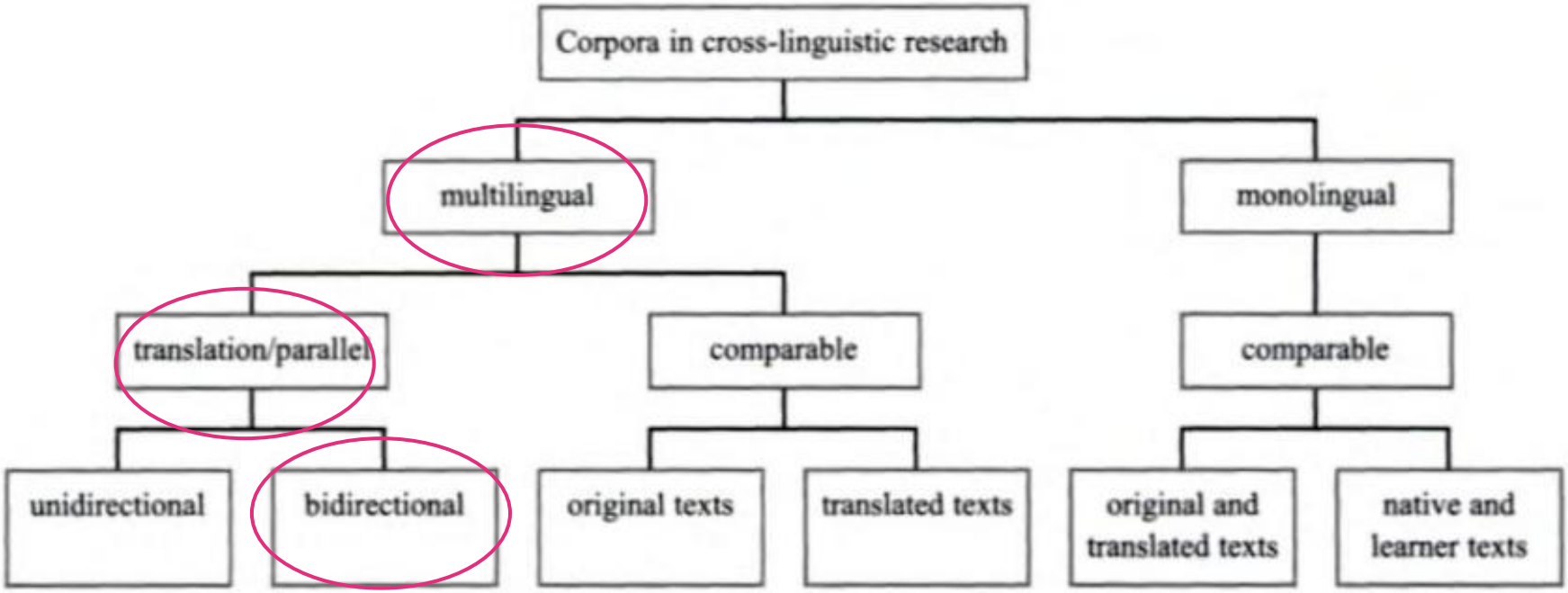
and its interlinear presentation of three languages

  - another inspiration:

Vulgate version of the Bible



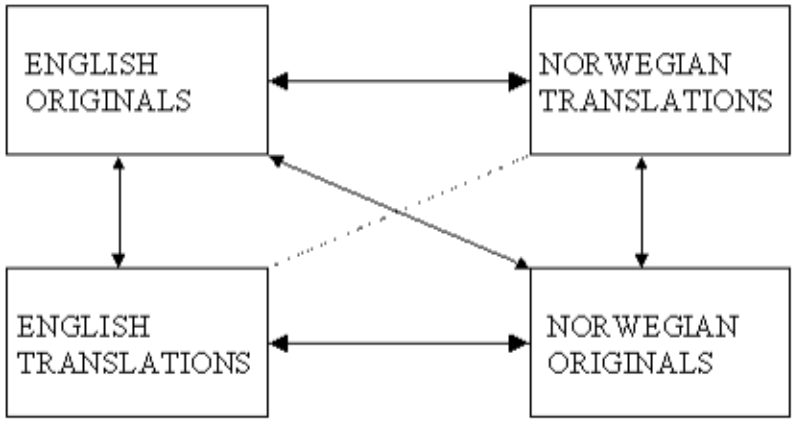
Corpora in TS/CS: terminology



See Granger S., Lerot J. & Petch-Tyson S. (2003) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam: Rodopi.*

Advantages of this parallel corpous

- both original and translated texts for comparison/reference



- *tertium comparationis*:
  - „background of sameness against which differences can be viewed and described“
- correspondence(s)
  - from source texts to translations, translator’s competence
  - between texts and languages
  - new insights on the languages compared

Current development

- research in corpus-based contrastive linguistics has recently ventured into new domains:
  - pragmatics & semantics
  - text linguistics
  - discourse
- increasing number of languages compared
- growing variety of topics and methodological approaches

- starting point: usually a preselected linguistic form or category with the aim to highlight similarities and differences in the structure, semantics or functions of the compared items across language boundaries, to reveal divergences in their use, the emergence of new meanings and language change

Corpus-based translation studies

The beginnings

- part of the descriptive translation studies branch (v. prescriptive)
- Toury, Hermans
  - to describe (i.e. explain) the specific characteristics of a translated text (or multiple translations of the same original) in terms of constraints or norms reigning in the target culture at a particular time that may have influenced the method of translating and the ensuing product.

- target-orientedness (Even-Zohar's theory of polysystem)
  - translated literature as a system worth of study in its own right
  - translated texts seen as specific and special
  - translations as a system in the target culture can be compared with non-translations in the target culture

Corpora in translation studies

- Mona Baker's seminal paper on CL and TS (1993)
  - the compilation of various types of corpora of both original and translated texts would enable translation scholars to uncover "*the nature of translated text as a mediated communicative event*"
  - the investigation of "universals" of translation, i.e. linguistic features that occur in translated texts and which are not influenced by the specific language pairs involved in the translation process
- translation universals (TU) v. source language effect (interference)
- TU not meant in a pejorative sense!

In search of a third code

- Frawley's term (1984): third code = the code (or language) that evolves during translation and in which the target text is expressed is unique
- not to confuse with translationese!, i.e. the unusual distribution of features that is clearly the result of the translator's inexperience or lack of competence in the target language

  "*translation results in the creation of a third code because it is a unique form of communication, not because it is a faulty, deviant or sub-standard form of communication*" (Baker 1993:248)

- translated texts record "*genuine communicative events and in this sense they are different from other communicative events in any language*". The nature of this difference, however, needs to be explored and recorded.

Translation universals

- Baker's original features of translation

1. simplification
   - the idea that translators subcosciously simplify the language or message or both
2. explicitation
   - the tendency to spell things out in translation, including in the simplest form the practice of adding background information
3. normalisation or conservatism
   - the tendency to conform to patterns and practices that are typical of the target language, even to the point of exaggerating them
4. levelling-out (convergence)
   - the tendency of translated text to gravitate around the centre of any continuum rather that move towards the fringes

Debate on the TU

- the concept of universals has been rather controversial
- general dissatisfaction with the Bakerian approach in last years
- many corpus-based translation studies have largely ignored the potentially important factors such as source language influence and genre variation
- vague definitions > difficult to operationalize
  - "*unmotivated, unparsimonious and vaguely formulated*" (Becher 2010)

> *"Research papers in the field should be minimally required to (i) provide a meticulous overview of the corpus materials used and of the exact procedures for selecting, annotating and sifting the data; (ii) comment on any specific problems encountered during data selection and annotation, including explicit and mo-tivated statements as to the solutions being adopted; (iii) include elaborate testing for statis-tical significance as a complement of, not in opposition to, thorough qualitative analysis."* (De Sutter et al. 2012)

Types of corpora in CTS

Different types of data

- Chesterman 2004: two kinds of research > two types of corpora

s-universals

interest in the comparison of translations with their originals (s = source texts)

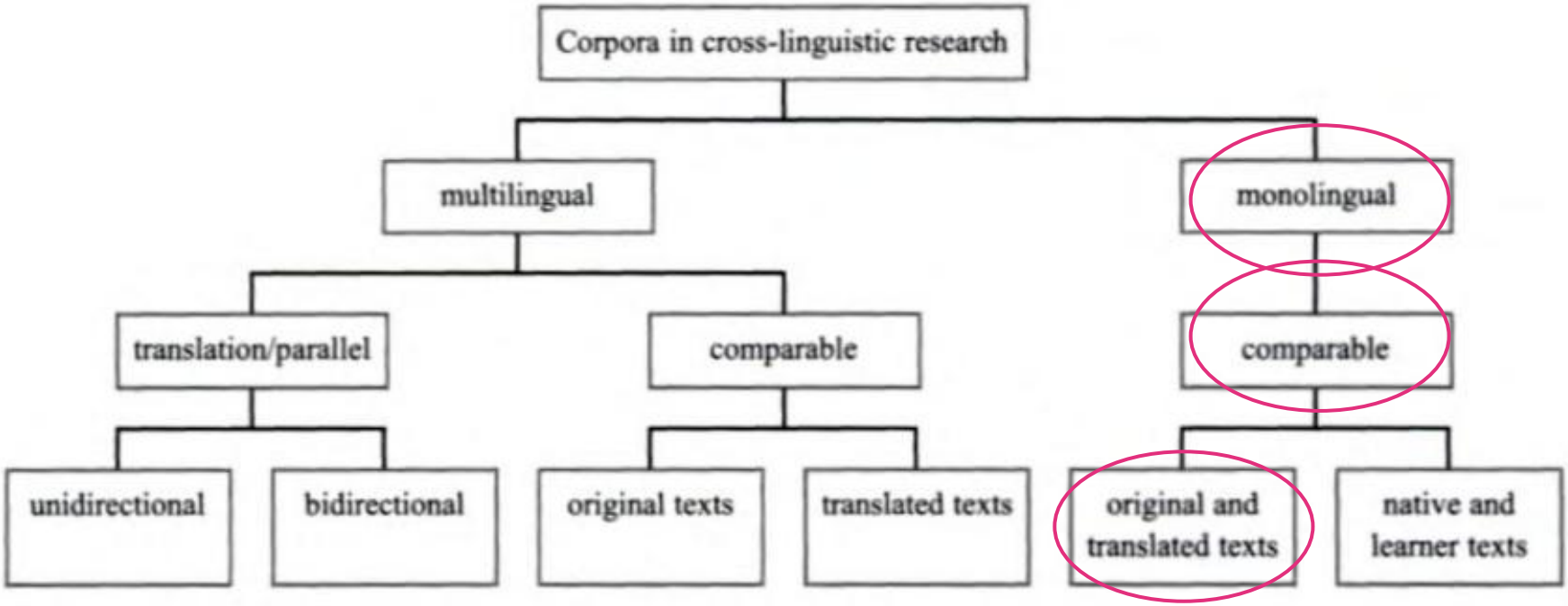
parallel corpus

t-universals

interest in the comparison of translations with non-translations (t = target texts)

monolingual comparable corpus

Corpora in TS/CS: terminology



See Granger S., Lerot J. & Petch-Tyson S. (2003) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies. Amsterdam: Rodopi.*

Monolingual comparable corpus

- usually designed to explore T-universals (features of translation analyzed against the non-translated language background)
- includes subcorpora of translations and non-translations compiled under the same (similar?) criteria:

SIZE

TEXT TYPE AND GENRE

DATE OF PUBLISHING

TEXT AND AUTHORS' HETEROGENEITY...

Obvious issues

- text size: translated and non-translated texts of similar genres may not be available in similar length
- disproportion in source languages: texts are not translated from different languages equally – English absolutely prevails
- nature of translated texts: translations from smaller languages tend to belong to „high-brow“ literature (Bernardini & Zanettin 2004), while translations from English etc. include any type of text
- cultural norms: translated texts may reflect different cultural/genre norms and therefore may not be directly comparable (e.g. cookbooks)

Case study: simplification in Czech

Simplification hypotheses

- Laviosa (1998: 8):

  Translated texts have a relatively lower percentage of content words versus grammatical words (i.e. their lexical density is lower).

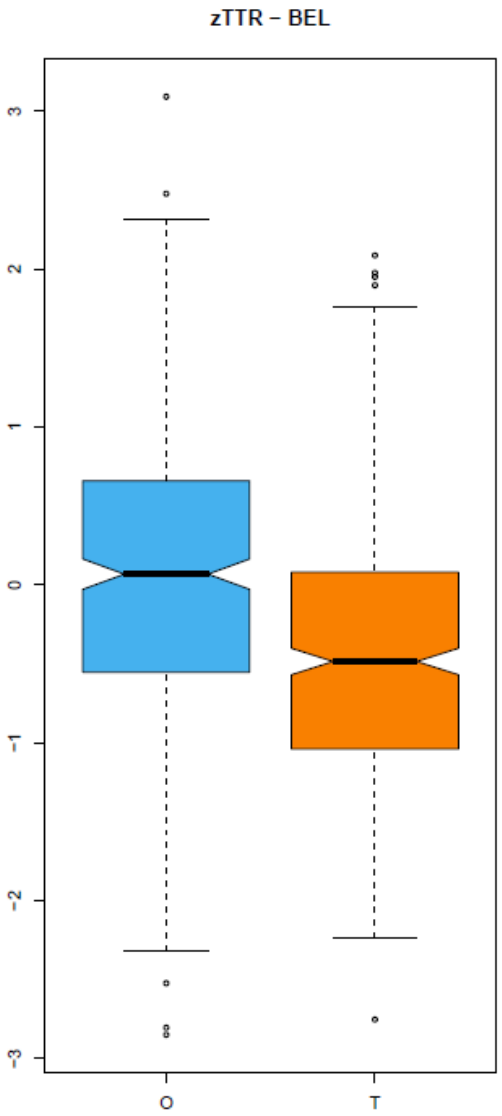
- Corpas Pastor & Mitkov & Pekar (2008):

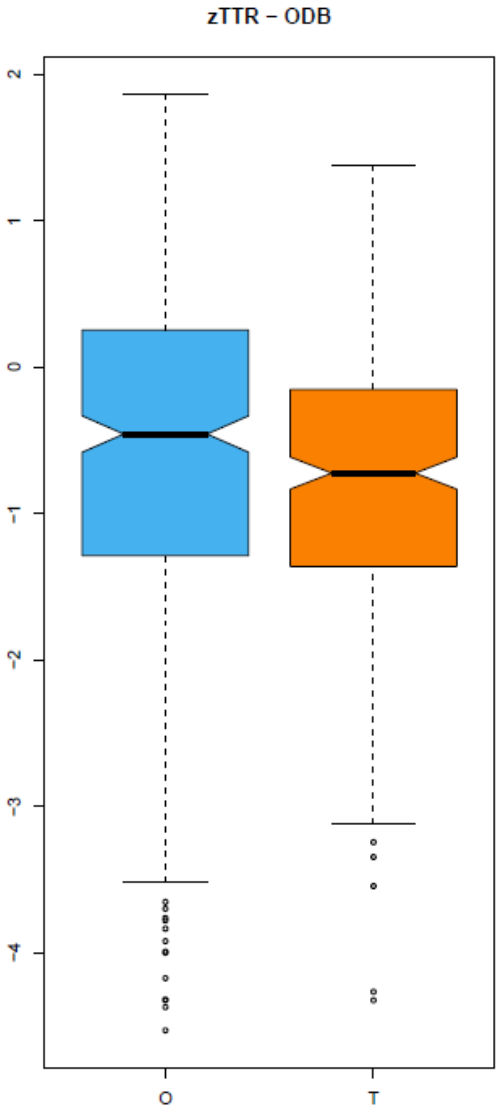
  We expect translated corpora to be characterised by less varied and more familiar vocabulary, […] to contain shorter sentences than sentences of original text.

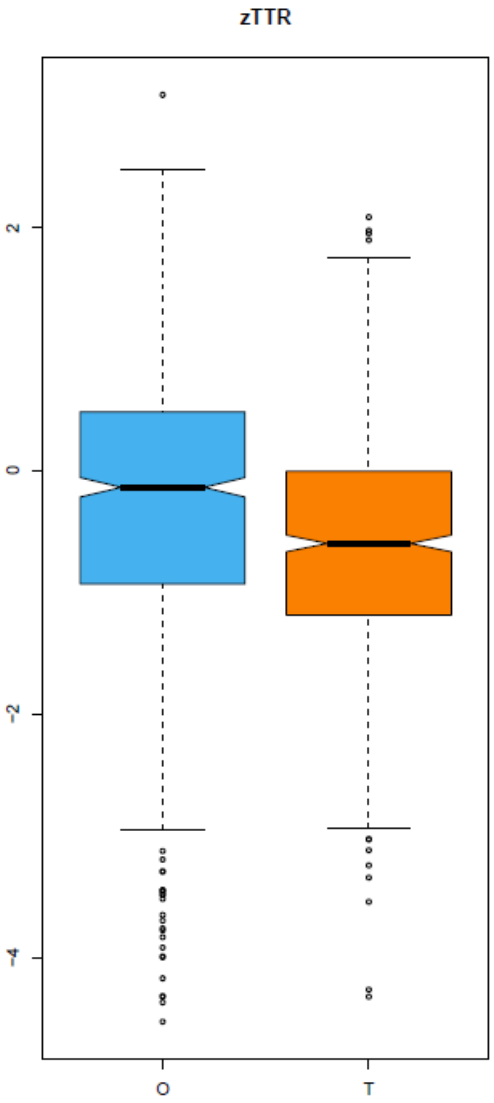
- Mihăilă (2010):

  The translated texts are said to contain a lower level of lexical richness and density.

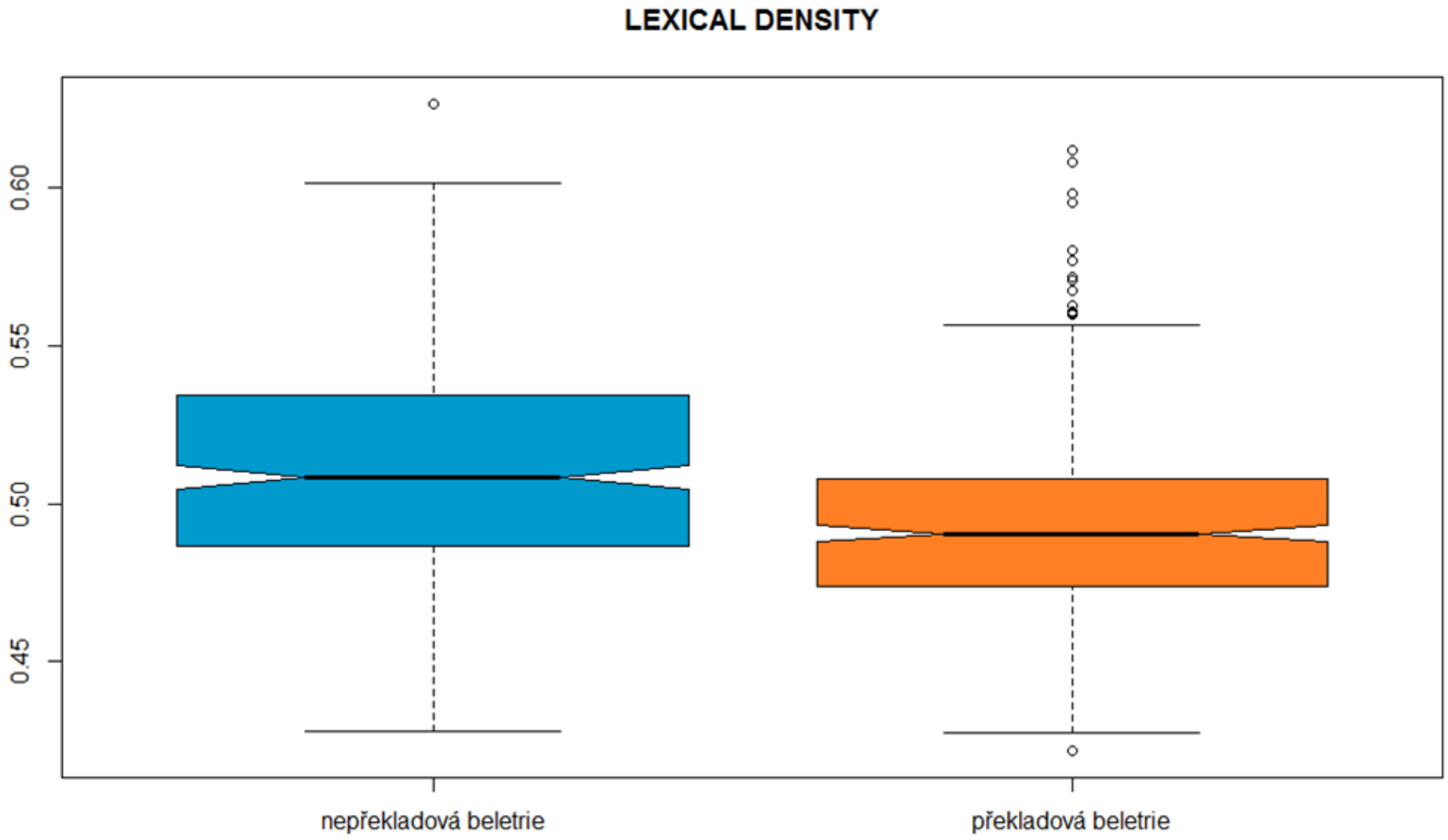
Simplification: TTR



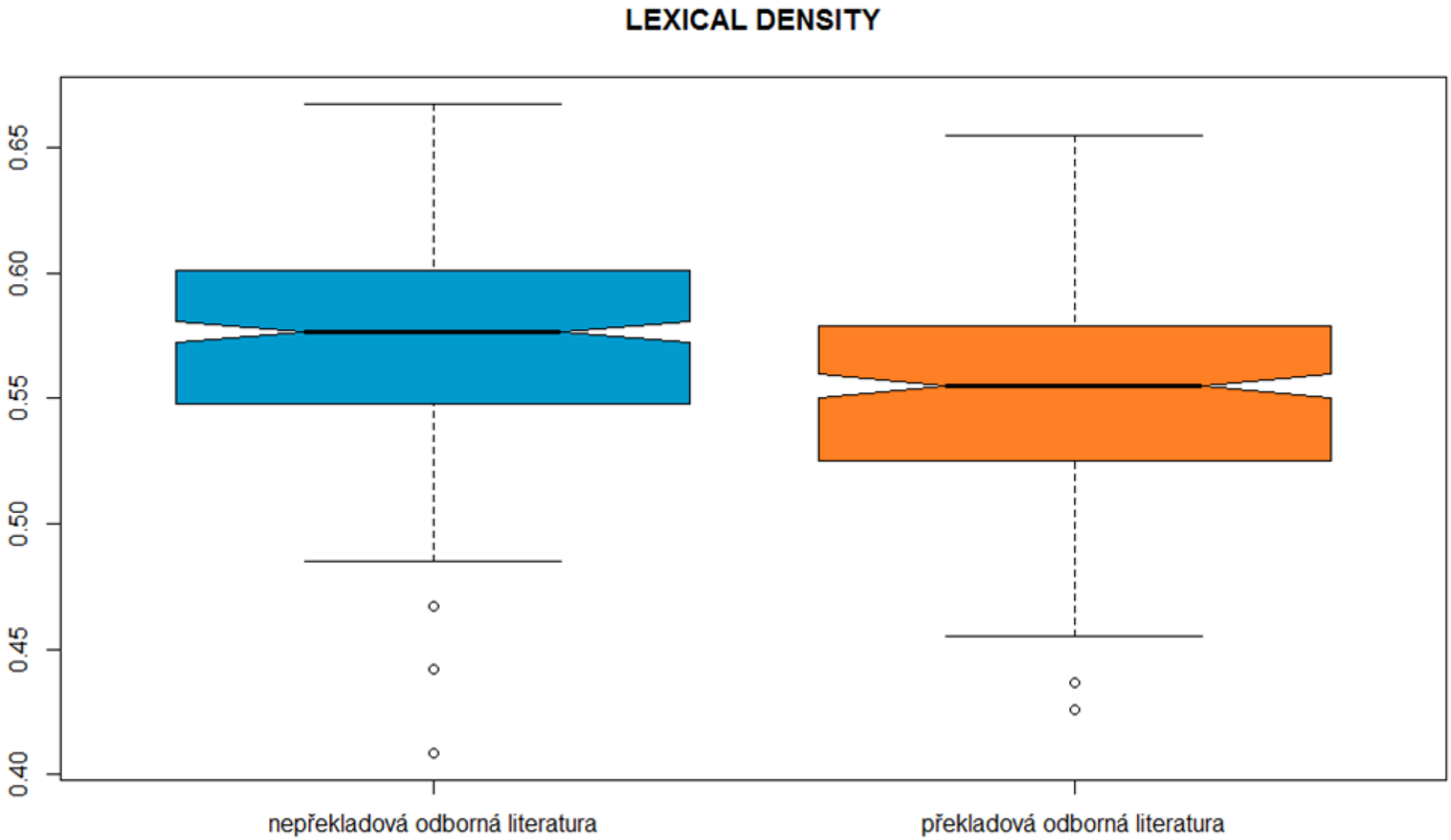




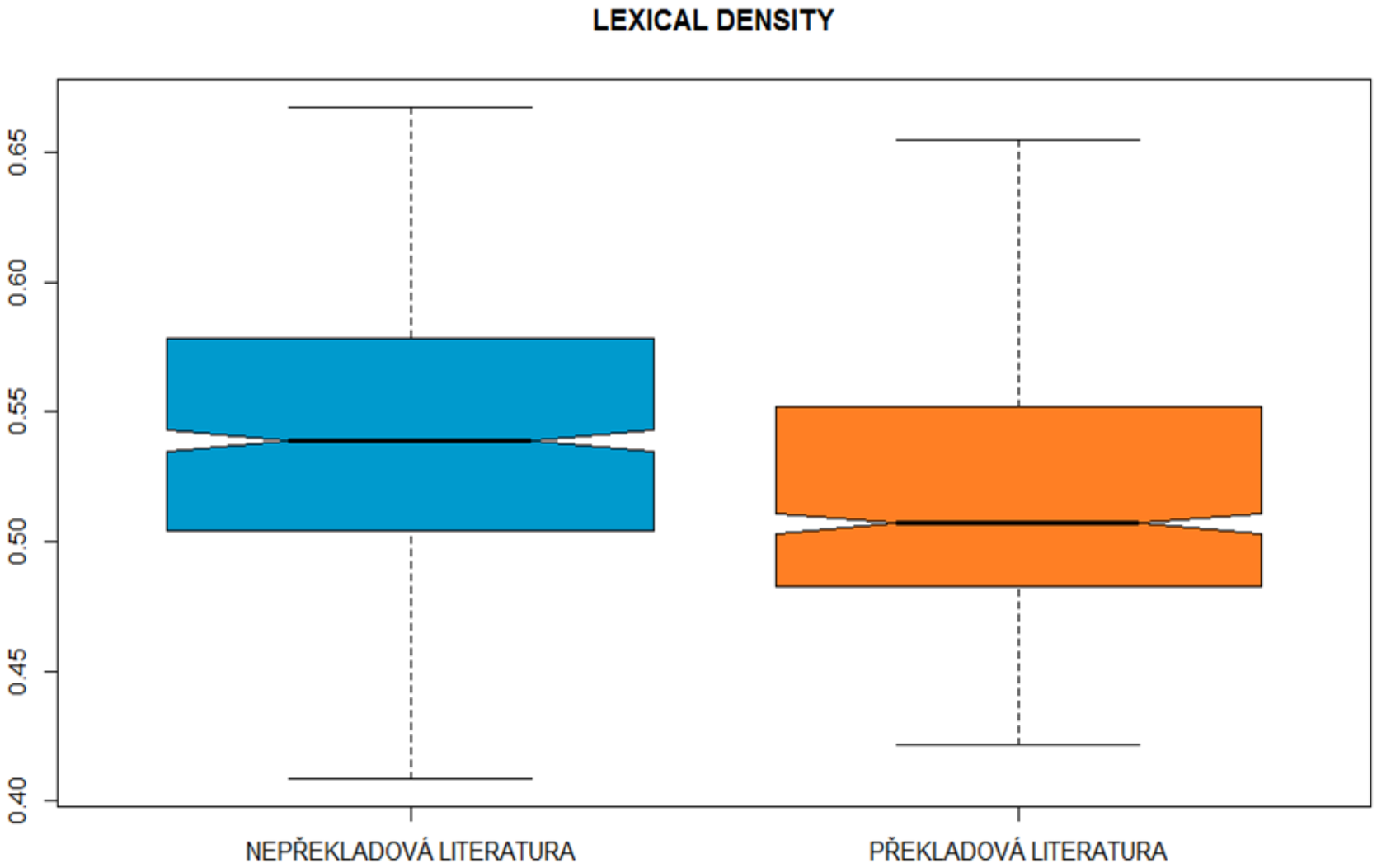
Simplification: LD (fiction)



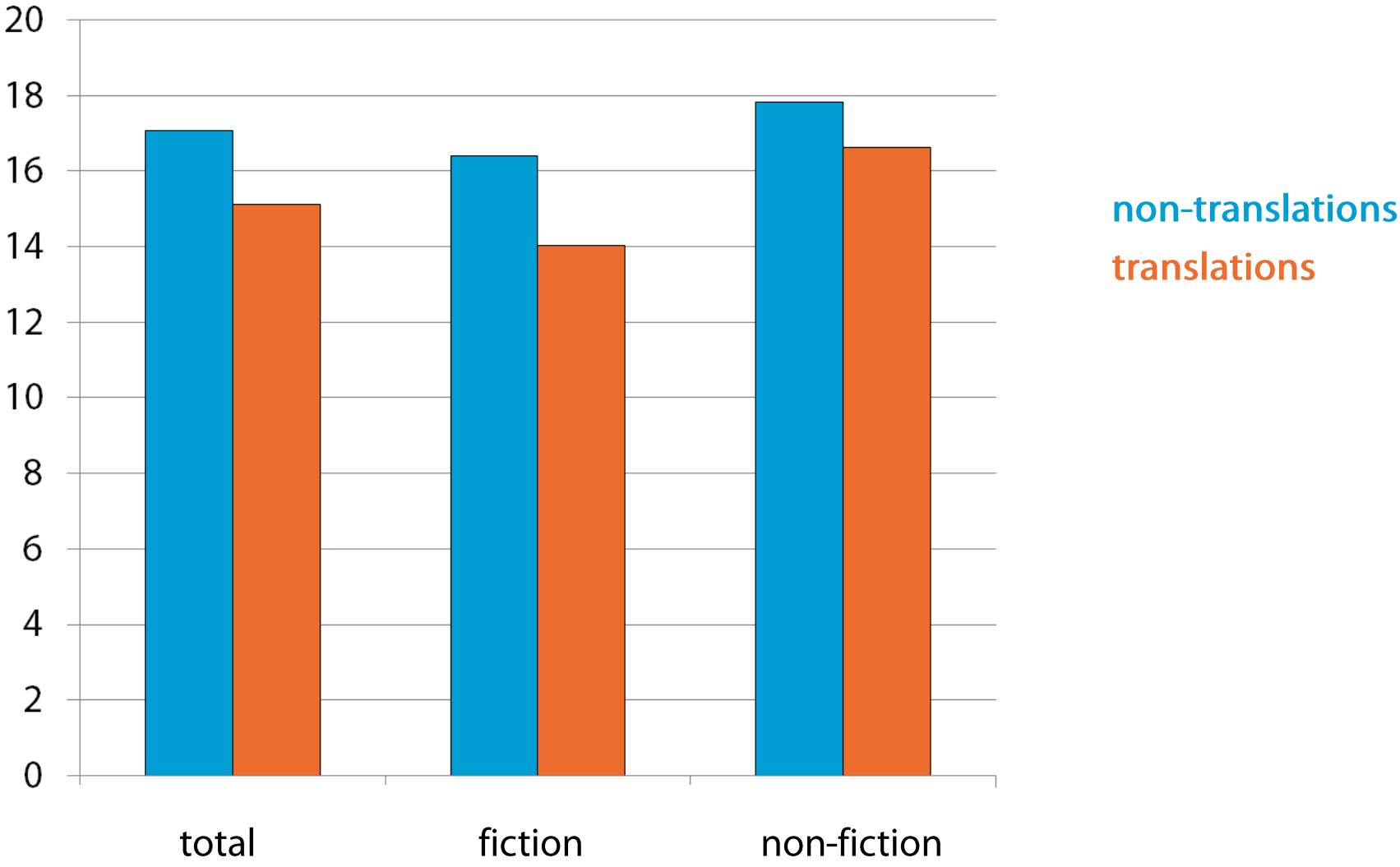
Simplification: LD (non-fiction)



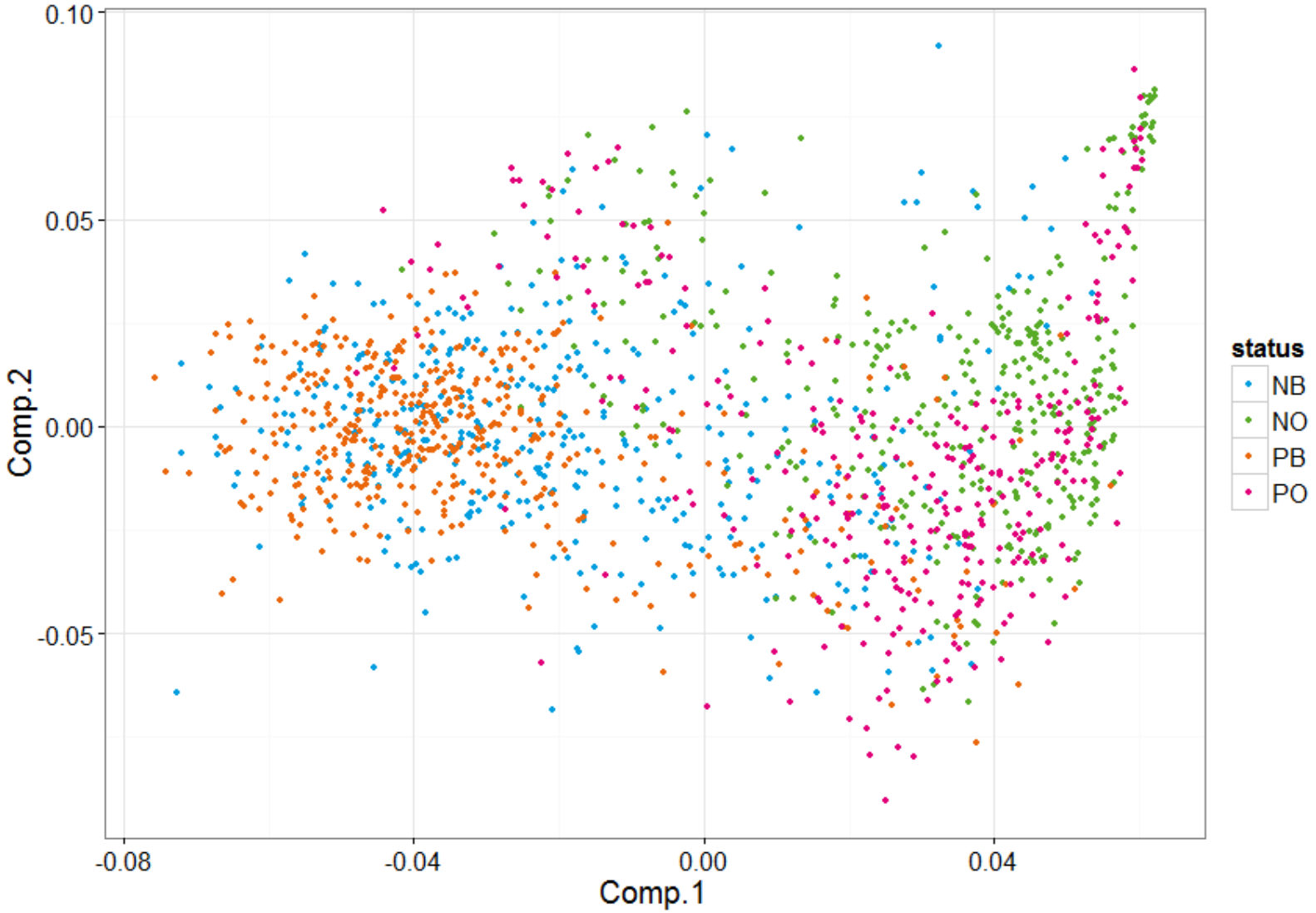
Simplification: LD (total)



Simplification: sentence length



Convergence: PCA (total)



Thank you for your attention!

Questions?

SEMINAR

Reading

common reading:

Chesterman, A. (20094). Hypotheses about translation universals. In G. Hanse, K. Malkmjaer & D. Gile (Eds.), *Claims, Changes and Challenges in Translation Studies*. *Selected Contributions from the EST Congress Copenhagen 2001.* (pp. 1–14). Amsterdam-Philadelphia: John Benjamins.

Discussion

- What is the difference between descriptive and prescriptive claims or hypotheses?
- What two main types of translation can we historically observe?
- What is the difference between s-universals and t-universals?
- How can they be tested?
- Should „bad“ translations be included in a parallel corpus?
- Can you think of any examples of features of translation in yout native language?