Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias

Christin Beck[1], Hannah Booth[1,2], Mennatallah El-Assady[1] & Miriam Butt[1]

[1]University of Konstanz, [2]Ghent University

The 14th Linguistic Annotation Workshop, co-located with COLING 2020, 12 December 2020

Motivation

- Development of linguistic corpora is fraught with **problems of annotation** and **representation**.
- Serious challenge for corpus designers and users.
- Undesirable consequences for research outcomes in NLP and theoretical linguistics.
- But so far not clear how to address the underlying problems within a **generally applicable framework**.

Related Work

- Existing approaches to representation problems:
  1. stochastic treatment (e.g. Dipper et al. 2013)
  2. assignment of ‘other’ category (e.g. Booth et al. 2020)
  3. left unannotated
- Some efforts towards more comprehensive schemes: Barteld et al. (2014); Lüdeling (2017); Merten & Seemann (2018); Pavlick & Kwiatkowski (2019).
- But **no generally applicable framework** as yet.
- **No understanding of how different problems interact** and potentiate in corpus development and use.

Our Paper

- We argue for a robust framework which explicitly treats representation problems.
- This paper represents a first step towards building a computational implementation for handling the underlying problems.
- Research is part of a larger effort on modelling representation problems in linguistic annotation processes via **visual analytics**.
- Conceptual basis: we identify and characterize **five sources of representation problems**.
- Extends discussion beyond ambiguity and uncertainty.
- Focus primarily on representation problems in **historical corpora**, but set of problems is transferable to other types of resources.

Identification: Five Sources of Representation Problems

Ambiguity: *one entity allows for multiple interpretations*

We propose **three categories**:

A Can be fully resolved
⟹ one interpretation

B Cannot be fully resolved, but identifiable preference:

(1) I saw some **cranes** by the river. The new apartments are starting to look really nice.

⟹ multiple interpretations, relatively ranked

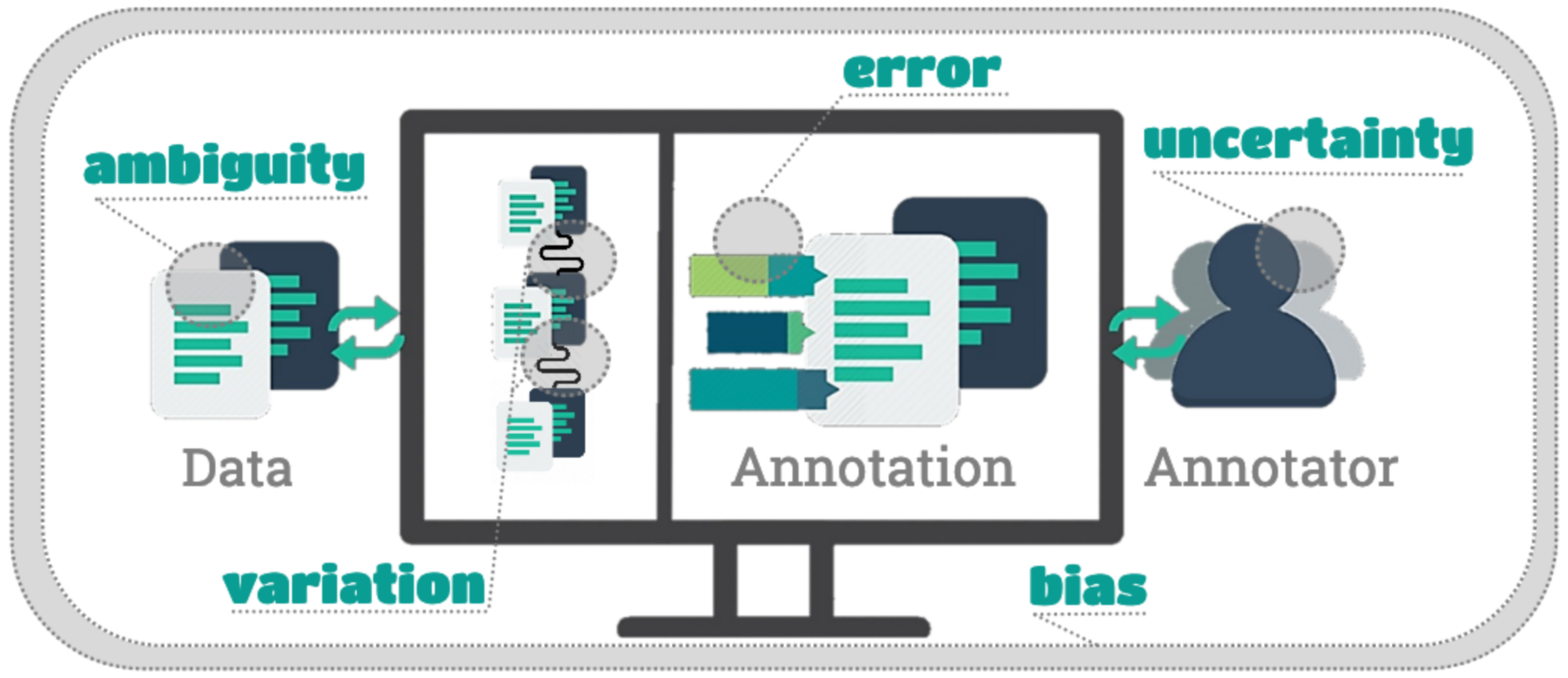
C Cannot be resolved, and no preference

(2) [His stupidly **missing** the penalty] lost us the game.

⟹ multiple interpretations, equal ranking

Error: *any representation which is not accurate with respect to the true value of an item*

Can be already **present in the data**, e.g. scribal errors, and/or introduced **in the preprocessing and annotation phases**



Variation: *a variable is expressed via multiple variants*

(3) a. Mary gave [an apple] [to John].
b. Mary gave [John] [an apple].

Our proposal: **link variants to a single variable**

Uncertainty: *multiple interpretations, but the relevant knowledge to opt for one is not available*

Esp. for **historical corpora**
⟹ Annotators lack native speaker competence and contextual knowledge

Bias: *an influence which leads to a preference or tendency for one thing over another*

Relevant for all phases of design and use, for instance:

- **Genre** bias
- Biased **NLP tools**
- **Theory** bias
- **Learning effect** during annotation

References

Fabian Barteld, Sarah Ihden, Ingrid Schröder & Heike Zinsmeister. 2014. Annotating descriptively incomplete language phenomena. In *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*, pages 99–104, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Hannah Booth, Anne Breitbarth, Aaron Ecay & Melissa Farasyn. 2020. A Penn-style Treebank of Middle Low German. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 766–775, Marseille, France, May. European Language Resources Association.

Stefanie Dipper, Karin Donhauser, Thomas Klein, Sonja Linde, Stefan Müller & Klaus-Peter Wegera. 2013. HiTS: ein Tagset für historische Sprachstufen des Deutschen. *Journal for Language Technology and Computational Linguistics*, 28:85–137.

Anke Lüdeling. 2017. Variationistische Korpusstudien. In Marek Konopka & Angelika Wöllstein, editors, *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*. IDS Jahrbuch 2016, pages 129–144. de Gruyter, Berlin.

Marie-Luis Merten & Nina Seemann. 2018. Analyzing constructional change: Linguistic annotation and sources of uncertainty. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality, TEEM’18*, page 819–825, New York, NY, USA. Association for Computing Machinery.

Ellie Pavlick & Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, March.

Challenges and Opportunities

Facilitating theoretical research

- Explicit treatment leads to a better understanding of the linguistic properties of the texts in a corpus.
- Advancing the respective state-of-the-art in theoretical linguistics.

Improving NLP models

- Propagating representation problems throughout NLP pipelines could inform computational models at each step.
- Improving the accuracy of algorithms and the resulting end-product.

Promoting reproducibility

- A generally applicable framework will avoid *ad hoc* treatments of representation problems.
- Solution to many barriers in the reproducibility crisis.

Guided annotation systems

- Such a framework could also inform guided annotation systems, which can adapt to annotators’ preferences over time and thus foster consistency and accuracy.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161. Many insights for the paper came from Hannah Booth’s work developing the Corpus of Historical Low German, with support from the Hercules Foundation/FWO, Grant number Hercules AUGE13/02 (2014–2015)/FWO G0F2614N (2016–present).