Data Science and Human Data: Annotation

Bonnie Webber

School of Informatics
University of Edinburgh
bonnie@inf.ed.ac.uk

November 24, 2014

Human annotation is valuable

Annotation involves marking up data:

- Identifying regions of interest in images or segments with particular properties in text;
- labelling those regions or segments.

In order to understand

- what makes those regions interesting or gives those segments their properties,
- what the labels follow from or correlate with,

human annotation is better than human intuition in helping to formulate a theoretically sound explanation or simply a reasonably accurate empirical model.

Human annotation is problematic

- People are often **inconsistent**: They say one thing at time t_1 and another at time t_2 [Klebanov & Beigman, 2009, 2010].
- People are often **biased**: They have preferences in how they answer questions and/or annotate data [Passonneau & Carpenter, 2014].
- Either can be the source of *inter-annotator disagreement*, but inter-annotator agreement (IAA) can mask both *inconsistency* and *bias*.
- It's neither efficient or sufficient to simply gather a huge amount of annotation of the same data: Crowd-sourcing alone is not the answer to either inconsistency or bias [Carpenter & Passonneau, 2014; Klebanov & Beigman, 2009, 2010].

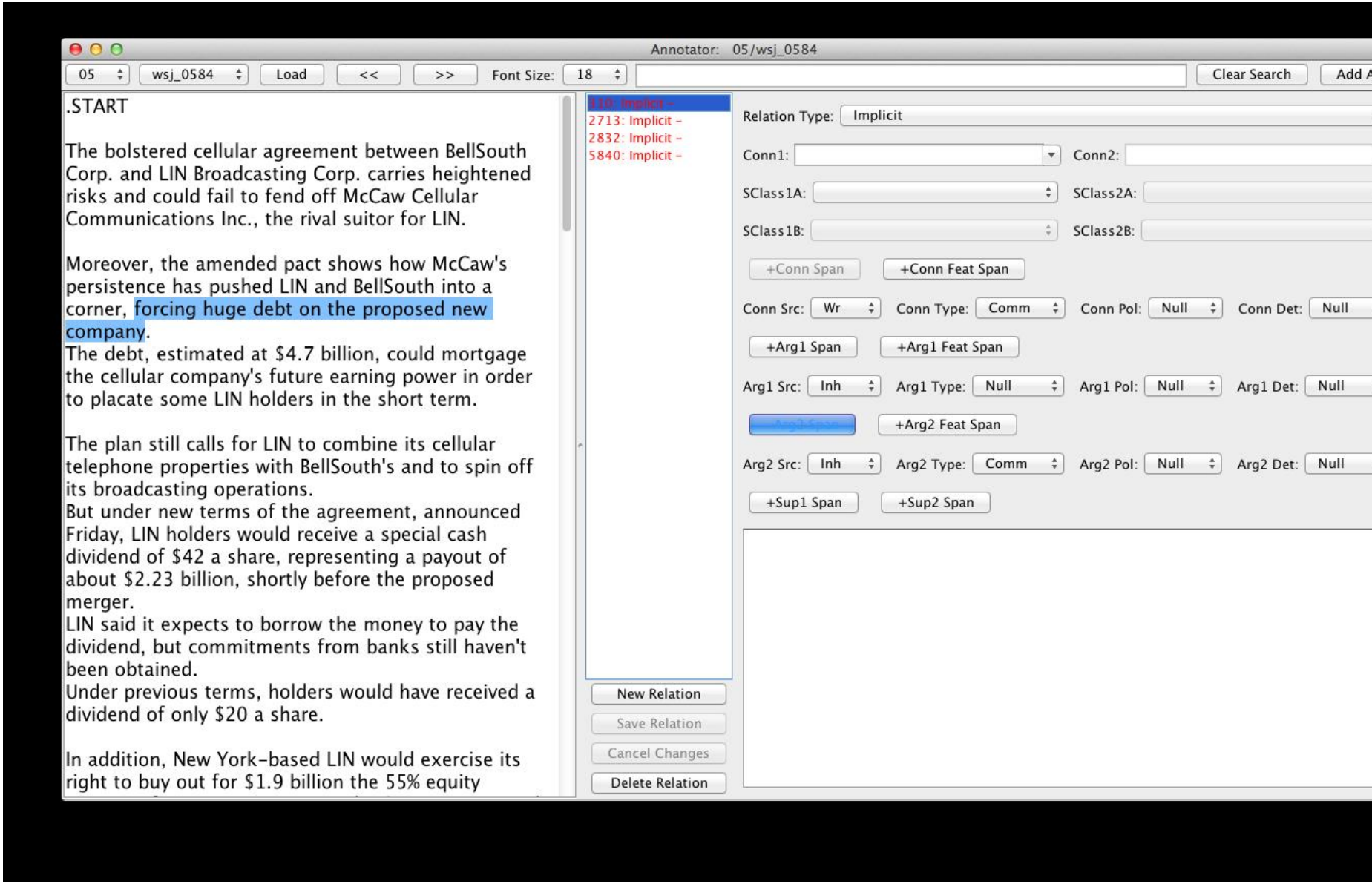
Tackling Inconsistency

- During annotation, one can try to show annotators their earlier annotatation of similar tokens.
- After annotation is complete, one can try to assess whether similar tokens have similar annotation.

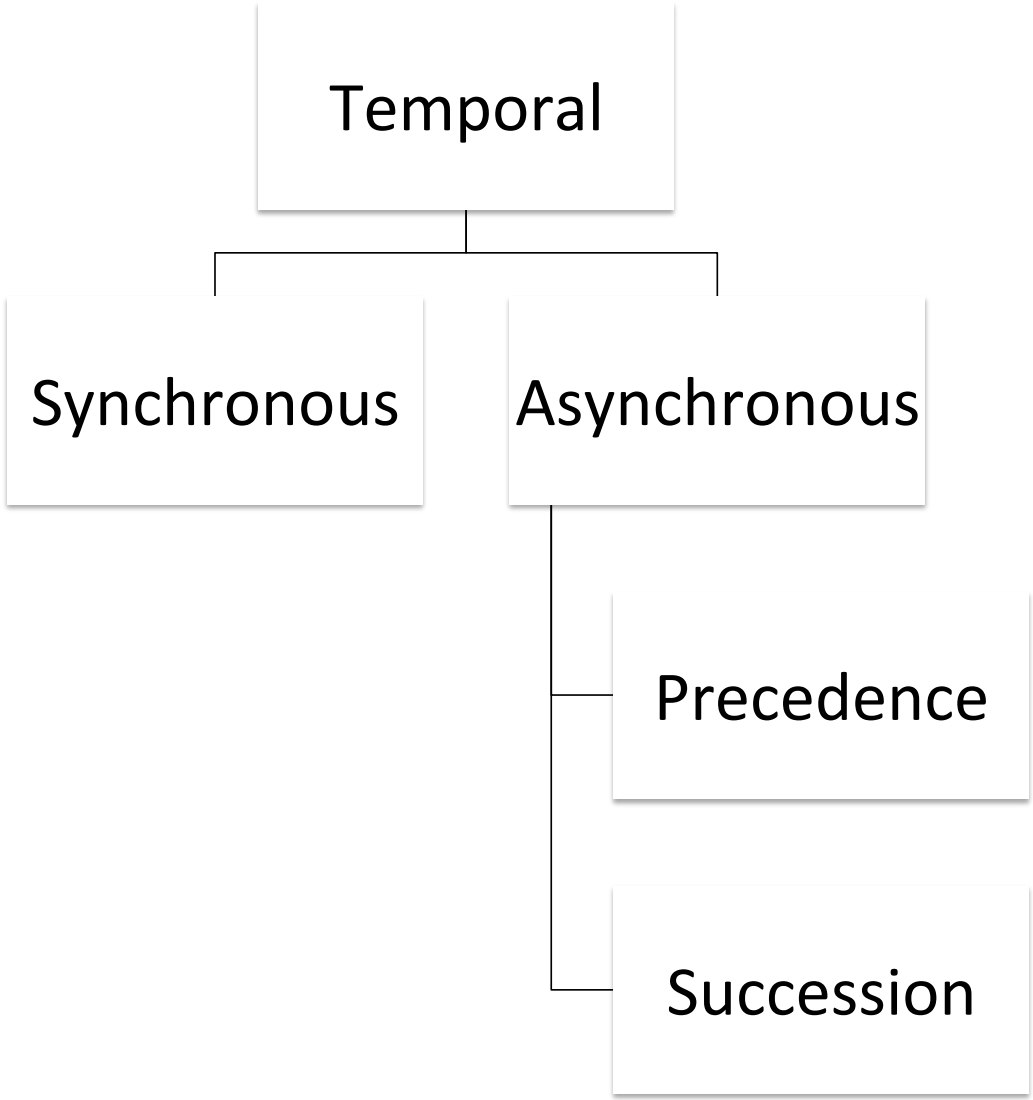
But both require the ability to identify *similar tokens*: For **discourse annotation**, this can be an interesting problem in its own right.

MSc project 1

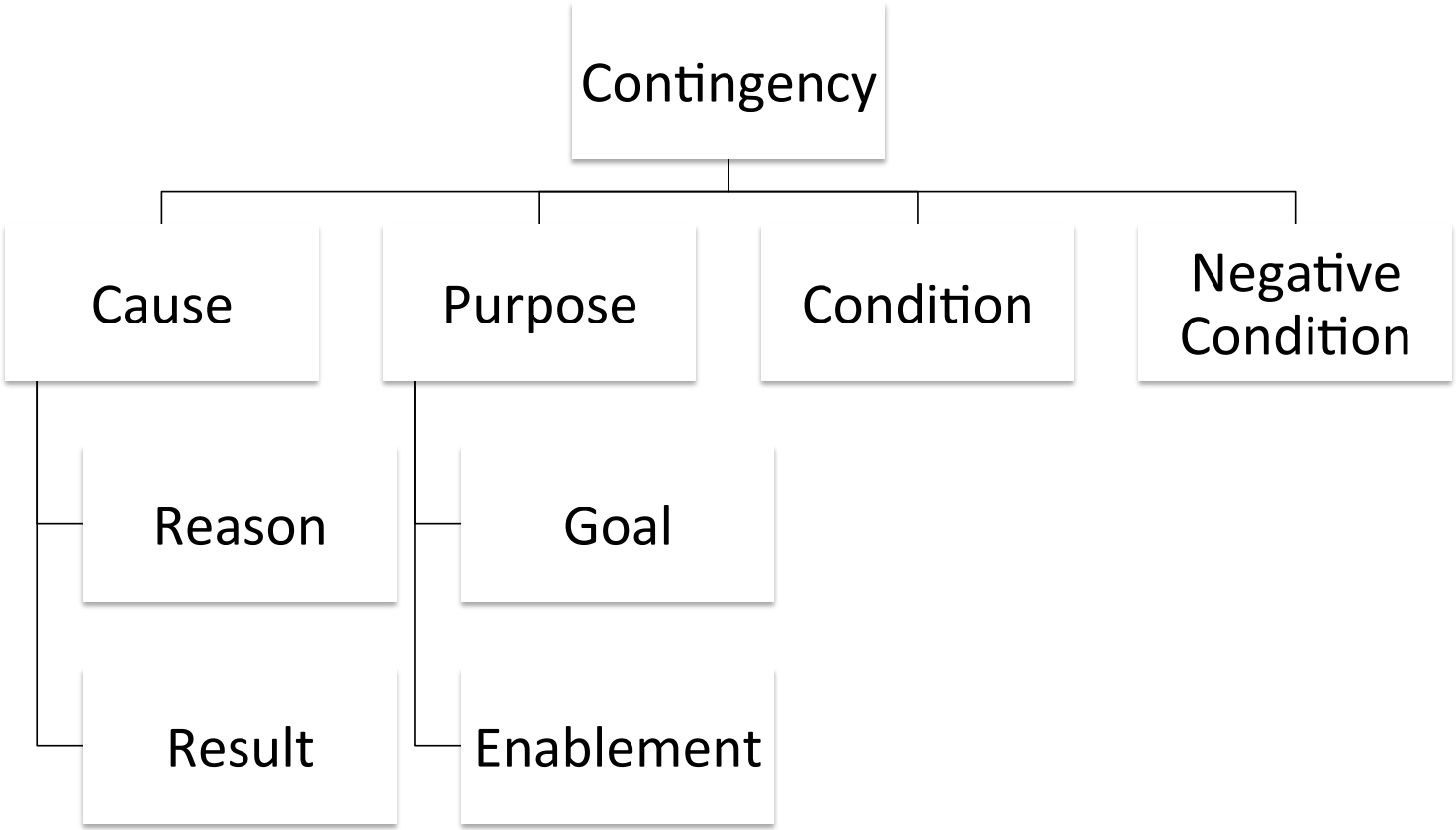
Consistency Project – Annotation Tool



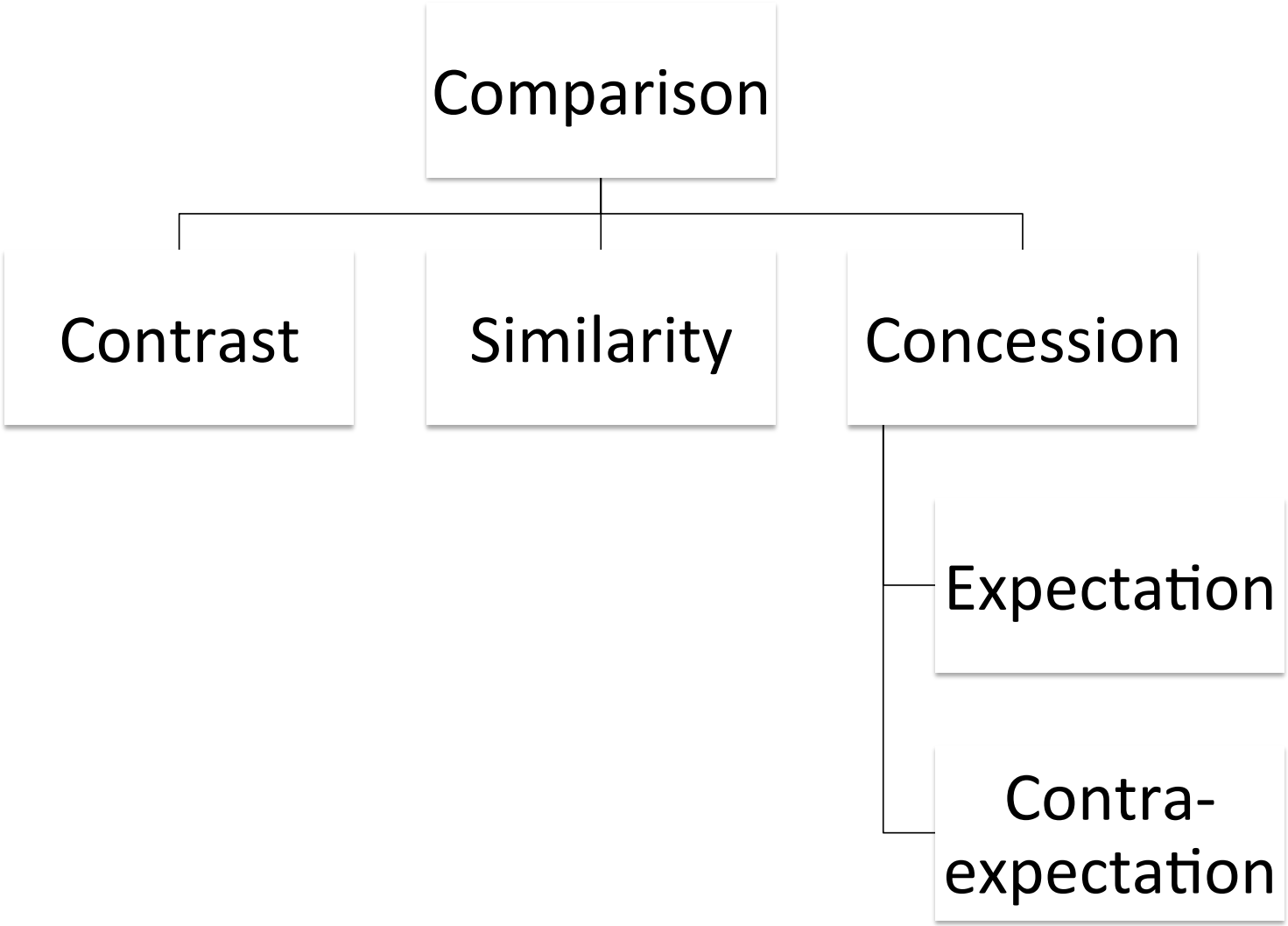
Consistency Project – Sense Annotation



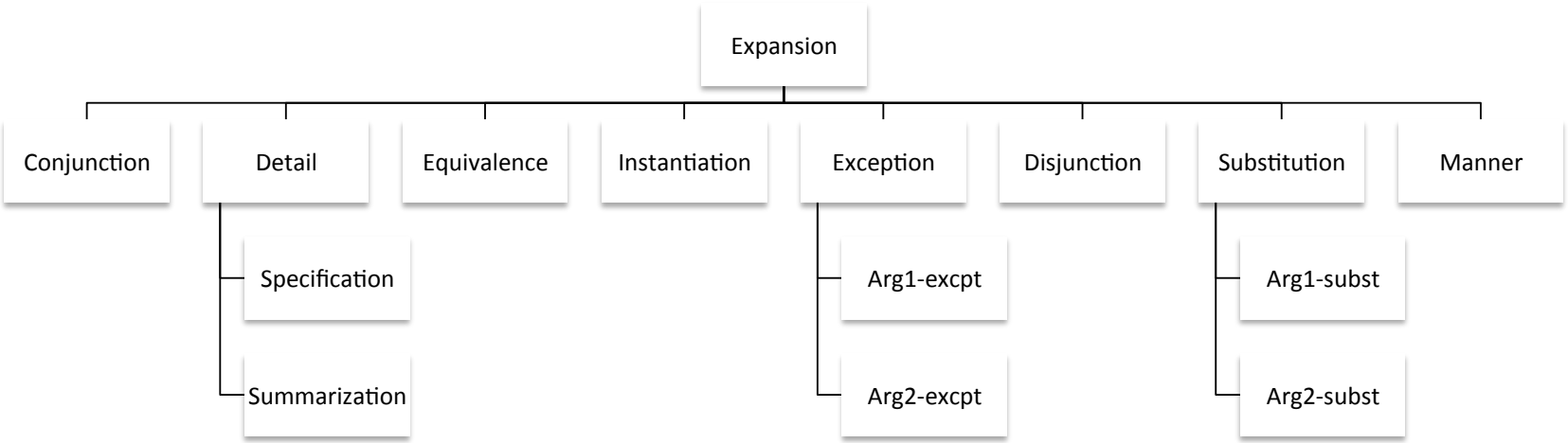
Consistency Project – Sense Annotation



Consistency Project – Sense Annotation



Consistency Project – Sense Annotation



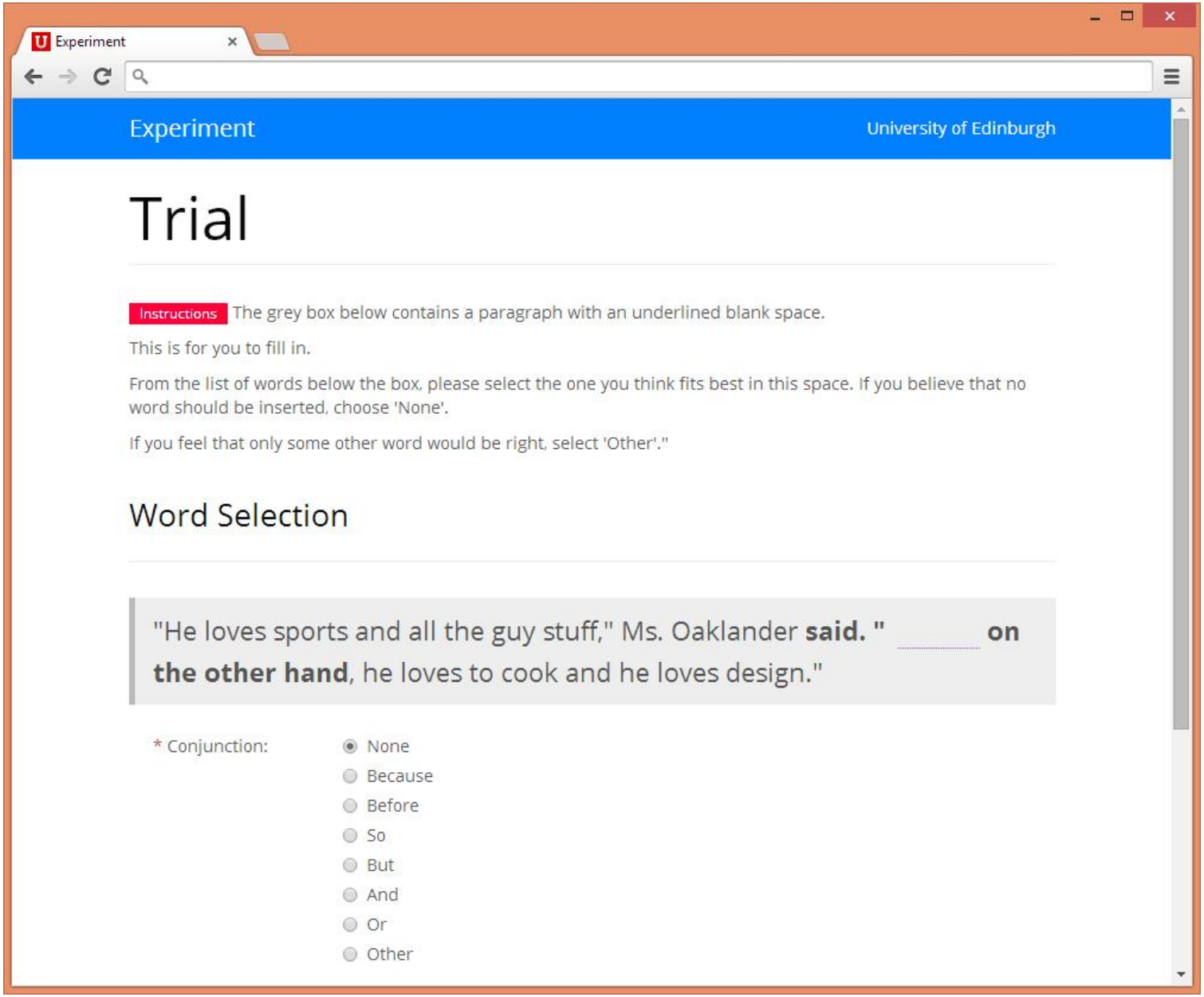
Tackling Bias

- Standard practice relies on inter-annotator agreement (IAA) to recognize biases that lead to annotators to assign different labels to the same token.
- Standard practice then relies on reconciliation to either reach agreement or make an executive decision.
- Neither ensures a high-quality corpus.
- Probabilistic models of agreement are more promising.

We'd like to experiment with this on **discourse annotation** that we are about to crowdsource.

MSc project 2

Crowdsourcing Experiment



References

◦ Beata Beigman Klebanov and Eyal Beigman (2010) Some Empirical Evidence for Annotation Noise in a Benchmarked Dataset *Proc. Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), pp. 438–446.

◦ Beata Beigman Klebanov and Eyal Beigman (2010) From Annotator Agreement to Noise Models *Computational Linguistics*, 35(4), pp. 495–503.

◦ Rebecca J. Passonneau and Bob Carpenter (2014). The Benefits of a Model of Annotation *Transactions of the Association for Computational Linguistics* (TACL), 2, pp. 311–326.